République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abdelhamid MEHRI - Constantine2
Faculté Des Nouvelles Technologies de l'Information et de la Communication
Département d'Informatique Fondamentale et ses Applications

N° d'ordre: . . /. . / . . . .

Série : . ./. . . / . . . .

THÈSE

Pour obtenir le diplôme de

Doctorat en Sciences

Spécialité : INFORMATIQUE

Présentée par :

Aicha AGGOUNE

Traitement de l'Hétérogénéité Sémantique pour l'Exploration des Sources de Données Multimédias

Soutenue publiquement le 23/05/2017 devant le Jury composé de :

Prof. Boufaida Mahmoud	Professeur à l'Université Abdelhamid Mehri Constantine 2	**Président**
Prof. Kholladi Mohamed-Khireddine	Professeur à l'Université Chahid Hamma Lakhdar d'El Oued	**Rapporteur**
Dr. Mezioud Chaker	MCA à l'Université, Abdelhamid Mehri Constantine 2	**Examinateur**
Dr. Derdour Makhlouf	MCA à l'Université Larbi Tébessi de Tébessa	**Examinateur**
Dr. Smaine Mazouzi	MCA à l'Université 20 Août 1955 de Skikda	**Examinateur**

Année Universitaire : *2016/2017*

Remerciements

Au nom d'Allah le tout miséricordieux, le très miséricordieux

قال الله تعالى: {وَإِذْ تَأَذَّنَ رَبُّكُمْ لَئِن شَكَرْتُمْ لَأَزِيدَنَّكُم}

Et lorsque votre Seigneur proclama: « Si vous êtes reconnaissants, très certainement J'augmenterai pour vous »

Je tiens à exprimer ma profonde gratitude envers mon directeur de thèse Mr. Mohamed-Khireddine KHOLLADI, Professeur à l'université d'El Oued, pour avoir accepté de prendre la direction de cette thèse en cours de route. Je le remercie pour ses conseils, son soutien continu tout au long de ma thèse. J'ai beaucoup apprécié ses qualités intellectuelles et humaines.

Je tiens à remercier très sincèrement Dr. Abdelkrim BOURAMOUL, Maître de conférences à l'université Abdelhamid Mehri Constantine 2, pour la confiance que vous m'avez témoignée et pour avoir bien voulu me proposer ce sujet de thèse ainsi que le degré de responsabilisation de son encadrement. Je le remercie pour ses remarques, ses orientations, ses encouragements, et sa rigueur scientifique qui m'ont permis d'accomplir ce travail.

Je remercie Mr. Mahmoud BOUFAIDA, Professeur à l'université Abdelhamid Mehri Constantine 2, pour avoir accepté d'être président du jury. J'ai également été très honoré par Dr. Chaker MEZIOUD, Maître de conférence à l'université Abdelhamid Mehri Constantine 2, Dr. Makhlouf DERDOUR, Maître de conférence à l'université de Tébessa et Dr. Smaine MAZOUZI, Maître de conférence à l'université de Skikda, qui ont accepté d'être examinateurs de ma thèse.

Je tiens à remercier toutes les personnes ayant participé et m'encouragé pour accomplir ce travail. Je remercie également tous ceux qui étaient vraiment heureux de mon succès et qu'ils me souhaitent toujours plus de succès dans tous les domaines.

كما قال الله تعالى: {وَقُل رَّبِّ زِدْنِى عِلْمًا}

Et dis: « Ô mon Seigneur, accroît mes connaissances. »

Résumé

Traiter le problème d'hétérogénéité sémantique pour l'exploration des sources de données multimédias est de nos jours une recherche de plus en plus demandée. Les approches d'intégration de données hétérogènes représentent une très bonne solution qui vise à offrir un accès unifié aux données sans besoin de connaître leurs sources d'origine. Cependant, ces approches sont rarement étudiées les données multimédias qui sont de nature complexe et difficile à représenter leur contenu sémantique. Ainsi, les sources de données multimédias peuvent être des bases de données multimédias ou des corpus des documents multimédias. Dans ce contexte, cette thèse présente la proposition de deux approches distinctes : la première est orientée base de données multimédias, en fournissant une médiation sémantique à base d'ontologies, et la deuxième est orientée document multimédia qui présente une indexation personnalisée à base du profil utilisateur.

Dans la première approche, l'exécution de requête d'utilisateur au niveau des adaptateurs consiste à appliquer le processus d'appariement entre la requête initiale exprimée en termes du vocabulaire de l'ontologie partagée ONTARIS (*ONTology of Alimentation RISks*) et les ontologies virtuelles propres à chaque base de données multimédia. Ce processus est basé sur quatre phases d'appariement en s'appuyant sur des mesures de similarité et le WordNet qui permet de désambiguïser le sens des mots. Nous développons également le système de médiation sémantique SAMER (*SemAntic Mediation for alimEntation Risks*) afin de réaliser nos expérimentations. La deuxième proposition présente une nouvelle approche d'indexation de documents multimédias dans le cadre du traitement du problème d'hétérogénéité sémantique. L'indexation personnalisée proposées est fondée sur l'exploitation des concepts communs entre le document et le profil utilisateur et les utiliser comme des index complémentaires aux index de base. Cette approche a été validée par la mise en œuvre de l'outil Persodexing (*Personalized indexing*). Une série d'expérimentations ont été effectuée et elle donne des résultats bien meilleurs.

Mots-clés: Appariement d'ontologies, Bases de données multimédias, Documents multimédias, Exploration de données, Hétérogénéité sémantique, Intégration de données, Indexation personnalisée, Médiation sémantique, Profil utilisateur.

Abstract

Dealing with the semantic problem for exploring multimedia data sources is nowadays a research more and more demanded. Heterogeneous data integration approaches represent a very good solution that aims to provide a unified access to data without the need to know their original sources. However, these approaches have rarely studied the multimedia data which are complex in nature and difficult to represent their semantic content. Thus, the multimedia data sources can be multimedia databases or corpuses of multimedia documents. In this context, this thesis presents the proposal of two distinct approaches: the first is an oriented multimedia database, by providing ontology-based semantic mediation, and the second is an oriented multimedia document which presents a user profile-based personalized indexing.

In the first approach, the execution of user query at the wrappers level involves the matching process between initial query expressed in terms of the vocabulary of shared ontology ONTARIS (ONTology of alimEntation RISks) and a virtual ontologies own to each multimedia database. This process is based on four matching phases using similarity measures and WordNet for word sense disambiguation. We are also developing the semantic mediation system SAMER (SemAntic Mediation for alimEntation Risks) in order to achieve our experiments. The second proposal presents a novel approach of multimedia document indexing for dealing with semantic heterogeneity problem. The proposed personalized indexing is based on exploiting the common concepts between the document and the user profile and using them as a complementary index to the basic index. This approach has been validated by the implementation of the Persodexing tool (Personalized indexing). A series of experiments have been carried out and it gives much better results.

Keywords: Ontology matching, Multimedia database, Multimedia documents, Data exploration, Semantic heterogeneity, Data integration, Personalized indexing, Semantic mediation, User profile.

ملـخـــص

معالـجة مشكلة عـدم التجـانس الدلالي لاسـتكشاف مصادر معطيـات الوسـائط المتعددة هو في الوقت الحاضر البحث الأكثر طلبًا . إن نُهج إدماج المعطيـات الغير متجـانسة تمثل حلًا جيدًا للغاية والتي تـهدف إلى توفـير وصول موحد للمعطيـات دون معـرفة مصادرها الأصلية . مع ذلك، هذه النُهج نادرًا ما تدرس معطيات الوسائط المتعددة التي هي ذو طبيـعة معقـدة وصعبة في تمثيـل محتـواها الدلالي . كـذلك، مصـادر معطيـات الوسـائط المتعددة يمكن أن تكـون قواعـد معطيـات الوسـائط المتعـددة أو مجـامع وثائـق الوسـائط المتعـددة . في هذا السيـاق، تقـدم هذه الأطـروحة اقتـراح نهـجين متـميزين: الأول خـاص بقـاعدة معطيـات الوسـائط المتعددة، و ذلك بتـوفير وسـاطة دلالية على أسـاس الأنطولوجيات، و الثاني خـاص بالوثائق الوسـائط المتعددة و ذلك بتقـديم فهـرسة مخصصة على أسـاس تعـريف المسـتخدم.

في النــهج الأول، تنـفيذ استعـلام المستخـدم على مستـوى المهيئـات يعتـمد على عمـلية المطـابقة بين الاسـتعلام الأولـي المبنـي على مفردات الانطولوجيا المشتركة ONTARIS (الأنطولوجيا في مخاطر التغذية) و الأنطولوجيات الافتراضية الخاصة بكل قـاعدة معطيـات الوسـائط المتعددة . تستند هذه العملية على أربع مراحل مطـابقة مستعمـلة معـايير التشـابه و WordNet الذي يسمح بإزالة الغموض في معنى الكلمات . كما قمنا بتطوير نظام الوساطة الدلالية SAMER (الوساطة الدلالية بالنسبة لمخاطر التغذية) لتحقيق تجاربنا. يعرض الاقتراح الثاني نهجا جديدا لفهرسة الوثائق الوسائط المتعددة في إطار معالجة مشكلة عدم التجانس الدلالي . تستند الفهرسة المخصصة المقترحة على استغلال المفاهيم المشتركة بين وثيقة و تعريف المستخدم و استخدامها كمؤشـرات مكملة للمؤشـرات الأساسية . و قد تـم اختبار هذا النهج بتطوير النظام Persodexing (فهرسة مخصصة). ولقـد أجـريت على كل المقترحات سلسـلة من الاختبارات التي أعطـت نتائج أفضـل بكثيـــر.

الكلـــمات المفتـــاحية: مطـابقة الأنـطولوجيات، قواعـد المعطـيات الوســائط المـتعددة، وثـائق الوســائط المتعـددة، استـكشاف البـيانات، عـدم التـجانس الدلالي، إدماج المعطيات، الفهرسة المخصصة، وساطة دلالية، تعريف المستخدم.

TABLE DES MATIERES

LISTE DES FIGURES

Liste des Figures

LISTE DES TABLES

LISTE DES ALGORITHMES

INTRODUCTION GENERALE

Ces dernières décennies, les utilisateurs sont devenus très intéressés par la production et l'exploration de données qui peuvent être privées (données personnelles, données d'études supérieures, compte Facebook, etc.) ou professionnelles fournies par différents secteurs que ce soit des données des entreprises, des laboratoires, des universités, des bancs ou encore du web (commerce électronique, E-Learning, etc.). De même, ces sources de données sont devenues plus hétérogènes et leur contenu est de nature multimédia qui ne permet pas de décrire facilement leur signification. Explorer ces sources de données multimédias produites des problèmes d'hétérogénéité sémantique liés non seulement à la diversité de définitions d'une même donnée mais aussi aux différentes interprétations et représentations de requête d'utilisateur. Explorer les sources de données multimédias avec une hétérogénéité sémantique réduite est de nos jours une recherche de plus en plus demandée.

1. Contexte du travail

Les données multimédias comportent plusieurs types de médias (textes, images, audios et vidéos) utilisés simultanément et stockées dans des supports informatiques décrivant les sources de données multimédias [Christodoulakis, 86]. Ces dernières peuvent être des bases de données multimédias ou des corpus de documents multimédias.

Dans le domaine de bases de données, les données multimédias sont représentées sous forme des enregistrements (ou n-uplets) des tables relationnelles (ou encore objet-relationnelles) ou une collection d'objets décrivant des classes d'une base de données orientée-objet. Quel que soit le type de modélisation des données multimédias (relationnelle ou orientée-objet), les bases de données sont appelées dans ce cas, bases de données multimédias ayant pour but de représenter ces données en utilisant des types complexes (les types LOBs, Image, Text, etc.) et d'effectuer des manipulations propres à chaque média à travers des requêtes multimédias (image requête, couleur similaire, etc.) disposant des opérations particulières permettant de gérer ce type de donnée (concaténation des chaînes de caractères, position d'un mot dans un texte, similarité entre descripteurs, etc.) [Stolze, 03].

En outre, les données multimédias peuvent être organisées dans des documents, appelés souvent documents multimédias. Un document multimédia décrit ses composants par diverses structures ou dimensions, les plus utilisées sont, la dimension logique, la dimension physique, la dimension spatiale, la dimension temporelle et la dimension sémantique [Laborie, 08]. L'ensemble de documents multimédias forme un corpus.

Explorer une source de données multimédias est donc traduite par une interrogation d'une base de données multimédia ou une recherche dans un corpus de documents multimédias. Ces sources de données sont par essence autonomes, distribuées et hétérogènes ce qui rend leur exploration un travail fastidieux.

Actuellement, le principal challenge scientifique est d'offrir un système capable d'explorer les sources de données multimédias sans aucun problème d'hétérogénéité sémantique qui induit à une dégradation des performances du système et une désorientation d'utilisateurs. Les travaux existants sur l'intégration de données visent à intégrer les sources de données hétérogènes de telle sorte qu'elles apparaissent comme une source unique [Lenzerini, 02]. De ce fait, un système d'intégration donne aux utilisateurs l'illusion de n'interagir qu'avec cette seule source [Wiederhold, 92]. Les travaux d'intégration existants dans la littérature consistent à fournir à l'utilisateur une vue réconciliée ou unifiée des sources de données. Plusieurs approches d'intégration ont été proposées, les plus utilisées sont :

- *Approche de médiateur :* elle présente une intégration virtuelle dans le sens où les données restent dans leurs sources d'origine et une interface unifiée du médiateur assure l'exploration de ces données via des requêtes exprimées en termes de vocabulaire du schéma global du système de médiation [Wiederhold, 92]. Le médiateur permet de réécrire la requête formulée en termes du schéma global en des plans de requêtes exprimées en termes de vue sur les sources de données. Des adaptateurs de chaque source de données doivent assurer l'exécution de ces requêtes réécrites après avoir les

traduire en un ensemble de requêtes compatibles au langage des sources. Les réponses obtenues dans chaque adaptateur seront combinées au niveau du médiateur en un seul résultat homogène et cohérent.

- *Approche d'entrepôt de données :* elle vise à créer une base de données unique appelée entrepôt de données (data warehouse) à partir des sources de données et de l'interroger par des requêtes exprimées en termes de vocabulaire du schéma de l'entrepôt [Inmon, 93]. La construction de l'entrepôt de données est basée sur le processus ETL (Extract, Transform et Load) qui permet dans un premier lieu d'extraire les données de sources via des extracteurs et d'appliquer diverses transformations aux données et dans un second lieu de charger les données transformées dans l'entrepôt [Vassiliadis, 03].

Une approche hybride qui combine à la fois l'approche de médiateur pour l'intégration des sources externes et l'approche d'entrepôt de données pour l'intégration de leurs données [Abiteboul, 02]. Dans ce contexte, cette thèse présente nos contributions de traitement du problème d'hétérogénéité sémantique pour l'exploration des sources de données multimédias.

2. Problématiques

La principale problématique est de définir un système capable de traiter l'hétérogénéité sémantique et d'explorer des sources de données multimédias. Dans le but de réduire le problème d'espace de stockage de données multimédias et de traiter l'hétérogénéité sémantique, nous nous intéressons à l'intégration sémantique par médiateur.

Beneventano et al [Beneventano, 13] ont proposé une approche étendue de médiation sémantique pour l'intégration à la fois des sources de données traditionnelles et multimédias. Cette approche est basée sur la fusion de deux systèmes existants : le système de médiation sémantique MOMIS utilise la logique de description ODL-I3 comme langage commun d'ontologie globale [Beneventano, 01] et le système de gestion de contenu multimédia MILOS [Amato, 04], pour assurer le stockage et la recherche par contenu de tous types de documents multimédias. L'approche étendue de médiation sémantique de données multimédias, bien qu'elle soit capable d'intégrer les données multimédias, il ne garantit pas une meilleure performance à cause de manque d'expérimentations dans des scénarios réels. D'une façon générale, il n'existe pas des travaux relatifs dans ce contexte et l'intégration sémantique de données multimédias plus précisément les bases de données multimédias reste l'un des verrous scientifique à lever.

Autour de cette principale problématique, plusieurs questions qui se posent les plus importantes sont :

- *Comment stocker et décrire la sémantique des données multimédias issues des bases de données multimédias?*
- *Comment déterminer et réduire le problème d'hétérogénéité sémantique?*
- *Comment assurer une médiation sémantique des sources de données multimédias?*
- *Comment réduire le problème d'hétérogénéité sémantique en recherche d'informations dans un corpus de documents multimédias?*

Cette thèse devrait présenter une nouvelle approche dédiée aux données multimédias dans le but de faciliter l'exploration et l'utilisation de ce type de données en traitant plus particulièrement le problème d'hétérogénéité sémantique. Dans la section suivante, nous présentons brièvement nos contributions pour répondre à ces différentes problématiques.

3. Contributions

Le travail présenté dans cette thèse vise à contribuer au traitement d'hétérogénéité sémantique pour l'exploration des sources de données multimédias qui peuvent être des bases de données multimédias ou un corpus des documents multimédias. Dans cette optique, ce travail porte sur deux principaux domaines de recherche : le domaine des bases de données (BD) et le domaine de la recherche d'informations (RI). Sur cette base, ce travail doctoral sera proposé deux approches distinctes : une approche de médiation sémantique pour l'exploration des bases de données multimédias et hétérogènes, et une approche d'indexation personnalisée de documents scientifiques et multimédias.

Notre travail porte d'abord sur une étude des caractéristiques des documents multimédias d'une part, et le problème d'hétérogénéité sémantique d'autre part [Aggoune, 12a]. En plus, nous pensons à améliorer la gestion des bases de données relationnelles pour être utiliser par la suite comme source de données multimédias [Aggoune, 12b]. Cette amélioration consiste à optimiser les requêtes en se basant sur la logique floue et la distance d'Hausdorff. Cette proposition a été raffinée dans [Aggoune, 15b] par la définition d'une nouvelle mesure de proximité sémantique entre requêtes à base de la logique floue.

Sur la base de cette première étude, nous orientons notre vision vers le domaine de la recherche d'informations dans le cadre de l'amélioration de recherche d'informations dans un corpus de documents traditionnels. Nous avons proposé une approche originale basée sur les requêtes géométriques où les termes d'une requête vont être vus comme un ensemble d'objets géométriques modélisés par le formalisme de la logique floue [Aggoune, 13c]. Les réponses de requêtes géométriques précédemment exécutées avec succès (ne contenant pas du bruit) seront utilisées comme réponses approximatives aux requêtes à réponses bruitées [Aggoune, 13b] [Aggoune, 13a].

En revanche, à cause de la complexité des données multimédias et de l'hétérogénéité des sources de données, nous avons trouvé des difficultés pour adapter les propositions précitées aux données multimédias. Pour cette raison, nous nous retrouvons devant le besoin d'apporter des solutions répondant au mieux à notre problématique principale.

Dans ce cadre, nous proposons dans un premier temps une approche d'indexation personnalisée de documents scientifiques et multimédias [Aggoune, 15a]. Cette approche étend l'indexation sémantique à base d'ontologie par l'ajout des nouveaux concepts complémentaires qui ont été identifiés à travers notre modèle d'indexation personnalisée de documents scientifiques et multimédias [Aggoune, 14a] [Aggoune, 14b]. Ce modèle a été obtenu par la mise en relation entre le modèle sémantique des documents multimédias et celui du profil utilisateur. Attribuer au document multimédia une représentation simplifiée et propre pour chaque utilisateur permet d'aider l'utilisateur à rechercher l'information désirée avec

moins des conflits sémantiques. L'approche proposée sera validée par la mise en œuvre de l'outil Persodexing (Personalized indexing).

Dans un second temps, nous proposons une approche de médiation sémantique à base d'ontologie permettant de traiter le problème d'hétérogénéité sémantique pour l'exploration des bases de données multimédias (BDMM) [Aggoune, 17]. L'approche proposée a la capacité d'intégrer à la fois des BDMM à base du modèle relationnel et du modèle orienté-objet. Cette approche s'appuie sur la création d'une ontologie partagée ONTARIS (*ONTology of Alimentation RISks*) de domaine de risques alimentaires, utilisée comme schéma global du médiateur et le processus d'appariement entre les requêtes d'utilisateur écrites en termes du vocabulaire d'ONTARIS et les ontologies virtuelles propres à chaque BDMM.

De ce fait, nous proposons des algorithmes permettant la création et l'évolution des ontologies virtuelles à partir des BDMM. Le traitement du problème d'hétérogénéité sémantique nécessite de faire des appariements ou matching au niveau de chaque adaptateur de base de données multimédia. Appliquer un appariement entre les éléments de requête et ceux des ontologies virtuelles, permet d'extraire toutes les relations sémantiques qui ne sont pas exploitées dans les travaux similaires. Le processus d'appariement se déroule en quatre phases fondamentales : appariement concepts, appariement instances, appariement propriétés et appariement relations. Ce processus s'appuie d'une part, sur le calcul des similarités sémantiques entre les pairs des éléments de requêtes et d'ontologie virtuelle et d'autre part, sur l'ontologie lexicale WordNet qui est essentielle pour désambiguïser le sens des mots. Le résultat d'appariement est un ensemble de correspondances qui seront transmises à la couche médiateur pour les rendre sous une forme normalisée à travers deux principales étapes : le prétraitement et la fusion. Le système de médiation sémantique SAMER (*SemAntic Mediation for alimEntation Risks*) devra être implémenté pour concrétiser et valider l'approche proposée. Ce système est facilement extensible en permettant l'ajout d'une nouvelle base de données multimédia avec une autocréation de son adaptateur et son ontologie virtuelle.

Nous finalisons notre travail par une série d'expérimentations pour évaluer les approches proposées. Les résultats de ces expérimentations montrent l'efficacité et l'adaptabilité de nos contributions pour le traitement du problème d'hétérogénéité sémantique. Ainsi, la comparaison avec des travaux similaires montre l'originalité de nos contributions. Sur la base des résultats d'expérimentations, nous déduisons quelques perspectives de recherche pour améliorer nos propositions.

4. Organisation de la thèse

Hormis l'introduction et la conclusion générale, cette thèse s'articule autour de cinq chapitres répartissent en deux parties principales. La première partie comporte trois chapitres sur l'état de l'art en rapport avec les thèmes qu'aborde notre travail. La seconde partie est composée de deux chapitres décrivant nos contributions.

Partie 01. Etat de l'art

Cette partie s'intéresse à l'élaboration de l'état de l'art selon trois champs de recherche liés à notre contexte : l'exploration de données multimédias, hétérogénéité sémantique et intégration de données, et la représentation de la sémantique via les ontologies. Ces trois champs sont également représentés par trois chapitres suivants :

Chapitre 01. « *Exploration de données multimédias* » : ce chapitre présente un état de l'art sur l'exploration de données multimédias. Pour cela, nous présentons en premier lieu, les notions de base liées aux données multimédias à savoir, la composition de données multimédias, les descripteurs liés à chaque média et les modes de représentation. Puis, nous présentons les stratégies d'exploration de données multimédias avec les problèmes liés à ce type de données.

Chapitre 02. « *Hétérogénéité sémantique et intégration à base de médiateur* » : En partant du dernier point du chapitre précédent, nous commençons à décrire en détail les problèmes des sources de données multimédias qui ont conduit à l'hétérogénéité sémantique lors de la recherche ou d'exploration. Par la suite, pour faire face aux problèmes d'hétérogénéité sémantique, l'intégration de données a pour but d'offrir un accès unifié aux données sans besoin de connaître leurs sources d'origine. De ce fait, nous présenterons deux principales approches qui ont été proposées dans la littérature: une approche virtuelle définie par le système de médiation et une approche matérialisée reconnaît par le système d'entrepôt de données. Dans cette thèse, nous nous focalisons sur la première approche et plus précisément la médiation sémantique de sources de données multimédias.

Chapitre 03. « *Ontologies : Un élément principal pour le traitement d'hétérogénéité sémantique*» : ce chapitre s'intéresse à la description d'ontologie qui joue un rôle crucial pour d'une part assurer une intégration sémantique et d'autre part de décrire la sémantique de données multimédias. Autour de cette ressource sémantique, nous présentons des approches, des méthodologies, des langages de représentation, de manipulation des ontologies et des techniques d'alignement. Ainsi, nous présentons un cas particulier d'ontologie linguistique WordNet que nous utilisons dans nos contributions.

Partie 02. Contributions

Cette partie aborde nos contributions et expérimentations effectuées pour atteindre nos objectifs. Elle est composée de deux chapitres ; le premier chapitre présente la première approche proposée orientée bases de données multimédias avec ses différentes expérimentations. Le deuxième chapitre expose la deuxième approche proposée orientée documents multimédias avec ses expérimentations et une analyse des différents résultats obtenus.

Chapitre 04. « *Approche proposée pour la médiation sémantique des BDMM : Présentation et expérimentations* » : ce chapitre expose notre première contribution qui présente une approche de médiation sémantique à base d'ontologie pour le traitement d'hétérogénéité sémantique lors de l'exploration des bases de données multimédias (BDMM). Nous présentons également les algorithmes nécessaires, les ressources utilisées et les mesures de

similarité sémantique appliquées. La deuxième partie de ce chapitre présente une série d'expérimentations avec une analyse des résultats obtenus.

Chapitre 05. « *Approche proposée pour l'indexation personnalisée de documents multimédias : Présentation et expérimentations*» : nous présentons la deuxième approche proposée à base du profil utilisateur pour l'indexation personnalisée de documents scientifiques et multimédias. Nous détaillerons les différents modèles sémantiques utilisés pour modéliser d'une part, les documents scientifiques et multimédias et d'autre part, le profil utilisateur. Ainsi, nous présentons le processus d'indexation personnalisée proposé. Une étude expérimentale doit être effectuée afin de montrer l'utilité et l'efficacité de l'approche proposée.

Cette thèse se termine par une conclusion générale qui résume les principales contributions et décrit des perspectives qui pourraient être envisagées un développement de recherche plus raffiné à partir de ce travail.

PARTIE I.

ETAT DE L'ART

Cette partie s'intéresse à l'élaboration de l'état de l'art selon trois champs de recherche liés à notre contexte du travail: l'exploration de données multimédias, hétérogénéité sémantique et intégration de données, et la représentation de la sémantique via les ontologies. Ces trois champs sont également représentés par trois chapitres.

CHAPITRE

01

EXPLORATION DE DONNÉES MULTIMÉDIAS

L'intitulé de cette thèse est le traitement de l'hétérogénéité sémantique pour l'exploration des sources de données multimédias. Avant de présenter les problèmes d'hétérogénéité sémantique, il devra nécessaire de connaître les différentes stratégies d'exploration ou de recherche des sources de données multimédias qui se diffèrent selon le mode de représentation. Ce chapitre a pour but de présenter les données multimédia, leurs composants, leurs descripteurs, leurs modes de représentation et les stratégies d'exploration. Il se termine par les différents problèmes liés aux données multimédias

1. Introduction

Actuellement, la majorité de données, qui circulent sur le web ou sont utilisées dans des systèmes d'information sont de nature multimédia. Les données multimédias comportent plusieurs types de médias (textes, images, audio et vidéo) utilisés conjointement. Cela, nécessite une représentation importante sur les différents descripteurs de média. Ces descripteurs peuvent être classifiés selon deux niveaux : les descripteurs du signal qualifiés de bas niveau et les descripteurs sémantiques qualifiés de haut niveau.

Dans ce premier chapitre, nous exposons les concepts de base liés aux données multimédias, leurs composants et leurs descripteurs. Ce type de données peut être représenté sous la forme de documents multimédias ou des n-uplets d'une base de données multimédia (BDMM). De ce fait, la recherche ou l'exploration des données multimédias peuvent être effectuées dans un corpus de documents via un système de recherche d'information ou dans une base de données multimédia grâce au système de gestion de base de données (SGBD).

Le reste de ce chapitre dresse l'état de l'art sur les modes de représentation et de recherche de données multimédias. Nous décrivons les types de larges objets (LOB) qui sont utilisés dans la plupart des SGBD pour définir les données multimédias. Nous présentons également nos exemples sous le SGBD Oracle qui est un très bon exemple pour définir et manipuler les BDMM. Enfin, nous en présentons les problèmes liés aux données multimédias afin d'apporter des solutions pour traiter ce genre de problèmes.

2. Données multimédias: définitions et composition

Pour définir ce qu'est une donnée multimédia, il est nécessaire de faire la différence entre les trois termes qui sont très proches entre eux: donnée, information et connaissance. Les données représentent l'information stockée dans un support de stockage pour être traitée, elles peuvent être des résultats de calcul, contenu d'une base de données, etc. [Boisot, 04]. Les données sont des informations non interprétées qui ne permettent pas de prendre une décision sur une tâche à réaliser. À l'opposé, l'information est une donnée ayant un sens permettant de prendre une décision sur une tâche donnée, tandis que la connaissance étend l'information par la capacité à mobiliser des informations pour agir [Boisot, 04].

En résumé, nous pouvons dire pour faire la différence entre ces trois termes:

- La donnée transporte l'information : ce sont des signaux non interprétés (information codée);
- L'information est une interprétation de la donnée ou le sens que l'on donne à celle-ci;
- La connaissance manipule l'information dans le cadre d'actions, dans un but précis, qui influence un processus. Les actions peuvent être la prise de décisions, la création de nouvelles informations, etc. Le passage de l'information à la connaissance est lié à l'expérience de l'action, cela veut dire veut qu'il n'a pas de frontière bien définie.

Le *multimédia* fait référence à l'utilisation simultanée de données de divers types tels que le texte, l'image, l'audio et la vidéo [Christodoulakis, 86]. L'étymologie du mot

multimédia vient de l'union de deux mots latins : *Multi* et *Média,* qui signifient littéralement plusieurs moyens de communication [Christodoulakis, 86]. Le terme média ou médium connote milieu, centre, intermédiaire ou encore un moyen de communication de la pensée [Christodoulakis, 86]. Il indique ainsi tous moyens de diffusion d'informations sous différentes formes (papier, radiophonique et télévisé) [Bulterman, 02]. Dans le domaine informatique, le terme média désigne un moyen de transmission, de stockage ou de présentation des informations [Bulterman, 02]. De telles données comportent plusieurs types de média utilisés conjointement sont appelées *données multimédias* [Yang, 12]. Ces données peuvent être rendues accessibles via le web ou stockées dans des bases de données multimédias. Nous limiterons notre travail aux données de type image et de texte qui sont représentées selon deux modes de représentations : les bases de données multimédias et les documents multimédias et scientifiques.

2.1. Le média Texte

Le média texte est le plus ancien, le plus significatif et le plus utilisé pour transmettre les informations par voie papier ou électronique. C'est le médium le plus existé sur le web dans des sources de données qui peuvent être des corpus de documents ou des bases de données. Le texte est un ensemble de chaînes de caractères alphabétiques ou même alphanumériques, bien organisé selon un certain nombre des paramètres de présentation (la police, la taille, le style, la couleur, etc.) et des paramètres de mise en page (alignement, marges, numérotation, etc.) [Jedidi, 05]. La représentation et le traitement du média texte sont assurés par plusieurs standards tels que RTF (Rich Text Format), PDF (Portable Document Format), DOC, LaTeX (Lamport TEX) [Mbarki, 08], etc. Le texte comprend trois aspects [Midouni, 16]:

- L'aspect lexical du texte correspond aux unités linguistiques telles que mots, verbes, etc.
- L'aspect syntaxique désigne les règles appliquées sur les unités textuelles.
- L'aspect sémantique concerne la signification des unités textuelles et les relations entre elles (synonymies, antonymies, polysémies, etc.)

L'interrogation de données textuelles s'intéresse à l'analyse des termes composant le texte selon deux aspects suivants [Jedidi, 05] :

- *La synonymie :* c'est la relation qu'entretient les divers mots ou expressions ayant le même sens ou un sens proche, par exemple : Aliment and Food.
- *La Polysémie :* c'est la propriété d'un mot qui présente plusieurs significations par rapport au contexte. Par exemple le mot "Book" n'a pas le même sens dans le contexte bibliographique (signifie un livre) que dans le contexte hôtelier (signifie réserver une chambre).

Dans ce cadre, la prise en compte du contexte dans l'interrogation de données textuelles constitue une voie prometteuse pour mieux répondre aux besoins en information de l'utilisateur [Bouramoul, 10]. Le média texte reste le média le plus représentatif du contenu sémantique de données que ce soit le texte lui-même ou les autres médias existants. Dans le

cadre de notre travail, nous ne pouvons pas présenter les données multimédias sans l'association du média texte pour présenter à la fois les documents multimédias et les bases de données multimédias. On verra dans les prochaines sections comment décrire ce type de média dans les deux modes de représentation de données multimédias (documents multimédias et bases de données multimédias).

2.2. Le média Image

Une synthèse de travaux de [Mbarki, 07] et [Jedidi, 05] nous a permis de définir le média image. Une image est une matrice composée de pixels et chaque pixel possède une couleur et éventuellement une transparence. Elle peut être présentée et codée selon plusieurs formats tels que : BMP, JPEG, TIFF, PNG, etc. Actuellement, les images représentent les informations en trois dimensions (3D), que ce soient des images photographiques ou générées par ordinateur. Une image 3D est obtenue par la liaison des formes désignant un objet réel tel que perçu par notre vision [Li, 07]. L'image ne porte aucune sémantique en elle-même contrairement au texte ainsi que son codage demande un temps assez long.

Le média image peut-être décrit par des trois types de paramètres:

- *Paramètres techniques:* ils correspondent aux, format du fichier image, sa taille et la représentation des couleurs.
- *Paramètres visuels:* ils correspondent à la couleur d'une image, la texture, la position, la forme, et l'orientation.
- *Paramètres sémantiques:* ils correspondent à l'annotation du contenu et de la sémantique à l'aide des métadonnées qui sont rajoutées aux images pour identifier, décrire et localiser les différentes ressources électroniques.

D'une manière générale, le média image est très utile pour illustrer rapidement des actions, des résultats, des personnages, etc. Néanmoins, utiliser une image seule sans annoter par des ressources textuelles ne permet pas d'expliquer les faits présentés dans cette image.

2.3. Le média Audio

Pour présenter le média audio, nous nous basons sur les travaux de [Lamel, 08] et [Rougui, 07]. Le média audio est représenté par des enregistrements sonores produits au moyen des matériels le plus souvent est le magnétophone. Ces enregistrements sont des signaux périodiques et continus caractérisés par la fréquence d'échantillonnage, la taille de l'échantillon, le pas de quantification, le nombre de canaux utilisé, etc. Le média audio est de nature temporelle décrite par trois facteurs principaux : l'instant de début de l'audio, la durée de lecture et l'instant de fin de l'audio. Une donnée sonore peut être présentée par plusieurs formats (wav, wma, mp, etc.) et sous différentes formes (parole, musique, bruit, etc.).

Le média audio est utilisé dans divers domaines les plus importants sont :

- La sécurité : par l'utilisation des alarmes indiquant les dangers ou l'existence des objets indésirables (ex. une arme).
- L'enseignement : pour apprendre une langue étrangère, illustration, etc. Il est donc nécessaire d'associer un casque pour écouter les documents sonores.

- La santé : l'audio est utilisée comme un élément fondamental d'aide pour les handicapés visuels.
- La télécommunication : pour établir une communication à distance entre personnes par l'utilisation de téléphone fixe ou mobile.

2.4. Le média Vidéo

Le média vidéo est le résultat de l'union et de la synchronisation de deux médias : image et audio [Spielmann, 10]. Une vidéo est un ensemble de séquences d'images forme des animations, muni des échantillons sonores qui apparaissent simultanément avec les images correspondantes [Spielmann, 10]. Spielmann affirme dans son livre que la vidéo est non seulement une étape intermédiaire entre l'analogique et le numérique, mais un média à part entière. Les formats de codage les plus connus pour représenter une vidéo sont : MPEG et AVI de Microsoft. Ainsi, la vidéo est le médium le plus volumineux qui peut être diffusé en flux (streaming) sur le réseau sous forme réduite en utilisant des outils de compression de vidéo qui permettent de diminuer l'espace mémoire occupé par la vidéo et de diminuer les débits de transmission [Spielmann, 10]. Actuellement, le média vidéo est le plus utilisé notamment pour la vidéoconférence, la vidéosurveillance, apprentissage, etc.

3. Descripteurs de données multimédias

La performance des systèmes de recherche et d'interrogation de données multimédias dépend pour une grande partie du choix des descripteurs utilisés pour décrire ces données. Le descripteur de données est une représentation simplifiée, consistante et intéressante qui décrit le contenu de données pour permettre de les retrouver aisément [Maron, 60]. Selon le type de média, nous distinguons quatre types de descripteurs : descripteurs de texte, descripteurs d'image, descripteurs d'audio et descripteurs de vidéo.

3.1. Descripteurs de texte

Une donnée textuelle est souvent considérée comme le type de données le plus représentatif du contenu sémantique. Elle est composée de paragraphes contenant des phrases successives qui composent des mots. La collection des mots les plus représentatifs du contenu de données peut représenter un descripteur [Moens, 06]. Ces mots sont généralement étiquetés par un poids représentant leurs degrés de représentativité du contenu sémantique [Moens, 06]. Dans la littérature, il existe plusieurs types de descripteurs textuels, les plus connus sont : la lemmatisation, la flexion et la radicalisation.

- *La lemmatisation* c'est l'opération qui prend en entrée l'ensemble de mots composant une donnée textuelle (ou factuelle) et donne en sortie un ensemble de lemmes [Gesmundo, 12]. Un lemme est une forme réduite et canonique d'un mot. Il peut être des verbes à l'infinitif ou autre mot (adjectif, nom, adverbe, etc.) au masculin singulier. Les lemmes peuvent également permettre d'associer des mots ayant une sémantique commune. En outre, toutes les entrées d'un dictionnaire sont des lemmes [Gesmundo, 12]. Par exemple l'adjectif long existe sous quatre formes : long, longue, longs et longues.

- *La flexion* représente les différentes formes d'un lemme caractérisées par des traits morphologiques qui peuvent être le genre, le nombre, le temps, etc [Moens, 06]. La construction d'une flexion à partir d'un lemme revient à ajouter des affixes (préfixes, suffixes) ou modifier le radical (racine) [Moens, 06]. Deux grandes catégories de flexions [Moens, 06]: la conjugaison pour les verbes et la déclinaison pour les autres mots (adjectifs et noms). La première catégorie, les verbes sont variés habituellement en personne, nombre, genre, temps et mode. La deuxième catégorie, les mots se changent de forme selon le genre, le cas ou le nombre.
- *La radicalisation* (en anglais stemming) est un procédé de transformation des flexions en leur radical ou racine (en anglais stem) [Ramasubramanian, 13]. Le radical d'un mot correspond à la partie du mot restante une fois que l'on a supprimé les affixes. Il ne correspond généralement pas à un mot réel contrairement au lemme qui correspond à un mot réel de la langue, par exemple, le mot argued (en français argumenté) a pour radical argu qui ne correspond pas à un mot réel.

Les descripteurs de média texte dépendent fortement la langue utilisée cela implique l'utilisation des techniques de TALN (Traitement Automatique du Langage Naturel) pour mieux décrire ce média. De plus, pour décrire correctement la signification d'un mot dans un texte, il est nécessaire d'associer des ressources linguistiques comme thésaurus, dictionnaire, ontologie, etc.

3.2. Descripteurs d'image

L'image est l'un des médias qui n'est pas capable de représenter explicitement son contenu sémantique, contrairement au texte qui est lui-même une source de description de la sémantique. L'image est caractérisée par des paramètres visuels qualifiés de bas niveau qui s'extraient via les techniques de traitement d'image et de signal [Smeulders, 00]. Ces paramètres ne permettent pas de décrire le contenu sémantique, ils sont souvent, la couleur, la texture et la forme. Dans cette optique, les descripteurs d'image ont été classifiés selon deux niveaux [Smeulders, 00]:

- *Descripteurs de bas niveau* : selon les paramètres visuels de l'image, il existe trois familles de descripteurs: descripteurs de couleur, descripteurs de texture et les descripteurs de forme.
- *Descripteurs de haut niveau* : décrivent le contenu sémantique de l'image. Ils dépendent d'une part des connaissances sur le domaine, exprimées à travers des ressources sémantiques telles que les ontologies.

L'image est composée d'un nombre limité de pixels [Anderton, 97]. Un pixel est le plus petit élément constitutif de l'image, il peut prendre des valeurs intermédiaires de gris ou de couleurs [Rani, 13] [Zhou, 10]. Le nombre de couleurs dépend le nombre des bits nécessaires pour coder un pixel, par exemple, pour un pixel codé en 24 bits, on peut y voir plus de 16 millions de couleurs. Les descripteurs de couleur sont généralement exprimés par des histogrammes qui représentent la distribution de couleur dans des espaces de couleurs tels que RGB (Red, Green et Blue), CMY (Cyan, Magenta et Yellow) et HSV (Hue, Saturation et

Value) [Eidenberger, 04]. Plusieurs mesures de similarité entre deux histogrammes de couleur ont été proposées dans la littérature, la mesure de base consiste à calculer l'intersection d'histogrammes de couleur [Swain, 91].

Les descripteurs de texture décrivent une image comme un arrangement spatial des pixels avec des variations locales de l'intensité lumineuse dans plusieurs directions et à différentes échelles [Picard, 95]. Ils sont basés habituellement sur des calculs statistiques sur les matrices de cooccurrences pour donner des indications sur le contraste, l'énergie, l'homogénéité, etc. [Zhou, 10]. Les matrices de cooccurrence définissent la probabilité jointe de l'occurrence de deux niveaux de gris quelconques dans une image [Bouguila, 07]. Puisque les textures sont la répétition d'un motif, des méthodes fréquentielles ont été utilisées pour décrire la texture telles que la transformée de Fourier, les filtres de Gabor et les ondelettes [Bouguila, 07].

Les descripteurs de forme permettent de présenter le contenu de l'image à partir de sa forme afin de décrire sa structure géométrique du contenu visuel [Eidenberger, 04]. Zhang et al., [Zhang, 04] ont proposé de classifier les descripteurs de forme en deux familles : *les descripteurs basés contour* qui décrivent les objets selon leur contour externe par l'application par exemple, de transformée de Fourier. *Les descripteurs basés région* représentent la distribution spatiale des pixels qui les constituent en utilisant des méthodes de calcul telles que l'enveloppe convexe, la surface, la compacité, etc. Ces descripteurs de bas niveau décrivent les images par leur contenu visuel.

Nous remarquons que les descripteurs de couleur sont robustes à certaines transformations géométriques de l'image. Toutefois, ils ne sont pas utiles s'ils sont utilisés seuls, car si deux images ayant la même couleur, il n'est pas forcément nécessaire qu'elles soient similaires, par exemple, une pomme rouge et une voiture rouge sont deux images possèdent la même couleur mais elles sont complètement différentes. Les descripteurs de texture sont très efficaces pour une recherche rapide et précise, mais ils ne donnent pas un meilleur résultat s'ils ont utilisé seuls. Enfin, les descripteurs de formes ne sont pas utiles s'il subit certains changements spatiaux d'une même image. Il est donc intéressant de combiner ces différents descripteurs pour une recherche plus efficace et plus discriminante.

Selon le papier présenté par [Hamadi, 15], les descripteurs d'images peuvent être appliqués dans la totalité de l'image, ils s'appellent dans ce cas les descripteurs globaux, ou plusieurs descripteurs locaux caractérisant chacun une partie de l'image. Un seul descripteur global décrit la totalité de l'image ce qui rend le temps de description très court, quoique robuste aux bruits, il ne permet pas de distinguer des parties de l'image contenant plus d'un objet [Zhang, 03]. Les descripteurs locaux où chaque descripteur décrit une partie de l'image cela assure l'invariance par translation contrairement aux descripteurs globaux. Les descripteurs locaux sont plus efficaces et ils permettent une recherche plus fine, néanmoins ils sont coûteux en termes de temps et du nombre de descripteurs utilisés pour une seule image [Wang, 06].

Tous ces descripteurs ne permettent pas d'exprimer le contenu sémantique de l'image. Le manque de concordance entre la représentation visuelle de l'image et sa signification est appelé le fossé sémantique [Smeulders, 00].

En effet, les descripteurs de haut niveau correspondent à des caractéristiques sémantiques. Il existe plusieurs méthodes de description du contenu sémantique de l'image. Le descripteur à base de texte enfoui dans l'image ou en surimpression, permet d'extraire des informations intéressantes du contenu de l'image [Clark, 02]. Il consiste d'abord à détecter l'existence du texte puis la localisation des zones textuelles dans l'image et par la suite l'extraction des informations [Clark, 02]. Ce descripteur est très efficace pour la navigation et la recherche des pages web contenant des images mais il n'est pas utile pour les images pures qui ne contiennent aucune zone textuelle.

D'autres descripteurs décrivent la sémantique des images par l'utilisation d'un banc de détecteurs d'objets résultant des classifieurs de concepts choisis manuellement et limités en nombre [Torresani, 10]. Ces descripteurs ont montré des performances intéressantes en classification et en recherche d'images, ils permettent de faire face à une grande variété de contenu. De plus, l'utilisation d'un plus grand nombre de détecteurs permet l'amélioration significative de la couverture conceptuelle et la gestion de plusieurs images [Li, 10]. Les descripteurs sémantiques à base des annotations consistent à associer des métadonnées sous forme des images clés ou de texte à une image donnée [Kustanowitz, 05]. Ces descripteurs sont de nature subjective ce qui génère le problème d'hétérogénéité d'annotation car deux annotateurs différents ne produiront pas systématiquement la même annotation pour une même image [Liu, 07]. Les annotations sont fiables mais dès que le volume de données soit grand, l'annotation devient une tâche longue et fastidieuse.

D'autres descripteurs sont basés sur les connaissances permettant de raisonner sur les représentations et les expressions du langage à travers des mécanismes d'inférence [Mbarki, 07]. Plusieurs modèles de représentation de connaissances, le choix du modèle dépend du niveau de complexité de données que nous utilisons.

En fait, l'équivalent de la sémantique dans les textes, c'est la sémiologie graphique dans les images ou les photos, elle permet de mettre un impressionnant pouvoir d'expression, qui nous est difficile de traduire en expression textuelle.

3.3. Descripteurs d'audio

Les descripteurs d'audio comme ceux de l'image sont classifiés selon deux niveaux : les descripteurs de bas niveau caractérisant l'ensemble de paramètres de signal et les descripteurs de haut niveau qui décrivent le contenu sémantique du son [Pinquier, 04]. A partir des changements de fréquence d'échantillonnage et de la segmentation temporelle du signal, on peut distinguer entre la parole, la musique et le bruit [Pinquier, 04]. Par conséquent, les descripteurs d'audio se diffèrent selon leurs types [Jedidi, 05]:

- Les descripteurs de parole sont extraits via les techniques de reconnaissance de la parole, à titre d'exemple le MFCC (Mel Frequency Cepstrum Coefficients) qui est très utilisé pour reconnaître les locuteurs à travers leurs voix [Ganchev, 05]. Le MFCC est

basé sur l'utilisation de deux paramètres de signal, les coefficients temporels et les coefficients de fréquences [Ganchev, 05]. D'autres descripteurs consistent à détecter les changements des locuteurs équivalents à des tours de parole dans un dialogue et de distinguer le genre de locuteur (homme, femme ou enfant) [Rougui, 07].

- Les descripteurs de musique sont nombreux tels que les descripteurs d'instruments et les descripteurs à base de pitch pour l'analyse et la synthèse de la musique [Peeters, 00].
- En ce qui concerne le bruit, un exemple de descripteur à base de volume et le taux de passage à zéro pour la détection de silence [Rougui, 07]. De plus, on peut y avoir des données sonores composant de ces trois types d'audio comme par exemple une parole sur musique avec des bruits. Dans ce cas, il suffit d'appliquer un descripteur propre à chaque type d'audio [Rougui, 07].

La description de contenu sémantique revient à associer aux données sonores des métadonnées de différents types [Pellegrino, 04]. Les métadonnées textuelles telles que les mots-clés d'un segment, le nom de locuteur, l'instrument, le titre du morceau, etc. Les métadonnées sonores qui sont exprimées par des sons clés ou des classes de sons telles que les classes explosion, cris, etc.

3.4. Descripteurs de vidéo

Pour la description de la vidéo, deux dimensions doivent être prises en compte : *la dimension temporelle* décrit la synchronisation des entités de la vidéo dans le temps et la *dimension spatiale* représente le placement de ces entités dans la vidéo [Jedidi, 05]. Ainsi, deux paramètres sont essentiels pour caractériser une séquence vidéo : l'environnement qu'elle représente et l'outil de son acquisition (caméra) [Jedidi, 05].

Les descripteurs de vidéo sont généralement définis à l'aide des métadonnées. Selon le nombre de médias composant une vidéo, nous pouvons distinguer des métadonnées des mots-clés extraites via la technique VOCR (Video Optical Character Recognition), des images clés et des sons clés qui représentent les objets (image ou son) les plus pertinents dans une séquence vidéo [Derbas, 14]. Des descripteurs du signal relatifs aux changements de l'environnement tels que les métadonnées liées aux changements d'illumination, de fonds et de scène d'une séquence vidéo [Sabri, 13]. D'autres descripteurs liés au capteur de vidéo incluent les paramètres essentiels de la caméra tels que le descripteur d'activité de mouvement (Motion Activity), la translation qui a été adoptée par le standard MPEG-7, la rotation et le zoom de caméra [Jedidi, 05].

Par ailleurs, la norme MPEG-7 (Motion Picture Engeneer Group) fournit une représentation standard des données multimédias [Jorgensen, 13]. Elle combine des métadonnées de bas niveau et de haut niveau à l'aide de langage XML (Extended Markup Language) pour supporter la recherche, l'édition, le filtrage et l'interopérabilité. Elle offre des outils de compression de données et elle permet également de décrire les relations temporelles et spatiales entre les objets qui composent une vidéo [Jorgensen, 13].

Le choix de descripteurs va permettre le stockage, l'indexation, la recherche et l'exploration efficace de données multimédias. Il est important que la description des données multimédias soit homogène selon l'application, le contexte et l'objectif à atteindre.

Dans le cadre de notre travail qui vise à traiter le problème d'hétérogénéité sémantique des sources de données multimédias et hétérogènes, nous nous focalisons seulement à la description sémantique à base d'ontologie de données contenant à la fois le texte et l'image.

4. Modes de représentations de données multimédias

Les données multimédias peuvent être représentées sous forme d'un document de différents formats de stockage ou des enregistrements (n-uplets) d'une base de données multimédia (BDMM). L'ensemble de documents multimédias sont regroupés dans un corpus de documents. Nous présentons dans cette section les deux notions principales pour la représentation de données multimédias : document multimédia et BDMM.

4.1. Document multimédia

Selon l'organisation internationale de normalisation ISO[1] (International Organization for Standardization), le concept de document se définit comme « *un ensemble constitué d'un support d'information et des données enregistrées sur celui-ci sous une forme en général permanente et lisible par l'homme ou par une machine* ». Dans un tel document, l'information véhiculée intègre différents types de médias (texte, images fixes ou animées, des graphiques, du son, de la vidéo) est appelé "*Document multimédia*" [Mbarki, 08]. Plusieurs standards pour la description de document multimédia, les plus importants sont [Mbarki, 08]: Dublin Core et MPEG-7.

Le Dublin Core est un schéma de métadonnées qui consiste à décrire le contenu (ex. titre, langage, etc.), les propriétés intellectuelles (ex. éditeur, droits, etc.) et l'instanciation de document multimédia (ex. format, date, etc.) [Weibel, 97]. Il permet d'améliorer la recherche de ressources complexes. Le MPEG-7 est un standard basé sur le langage XML, il vise à décrire le document multimédia par un ensemble de descripteurs [Manjunath, 02]. Ainsi, il facilite la recherche et l'indexation de documents multimédias à travers la compression de données.

L'organisation de différents médias composant un document multimédia se fait selon plusieurs dimensions, les plus utilisées sont la dimension logique, la dimension physique, la dimension spatiale, la dimension temporelle et la dimension sémantique [Laborie, 08].

Le reste de cette section présente ces quatre dimensions du document multimédia qui décrit la cascade thermale de Guelma en Algérie, illustré dans la figure 1.1. Ce document a été construit en se basant sur le contenu du site web :

http://www.yasminetravel-dz.com/content/hammam-maskhoutine

[1] http://www.iso.org/iso/home.html

Figure 1.1. Exemple du document multimédia "Cascade thermale"

4.1.1. Dimension logique

La dimension logique de document multimédia repose sur la description hiérarchique des entités composant le document et éventuellement les relations existantes entre elles [Laborie, 08]. Le langage XML est un exemple de langage de description de document multimédia qui permet de représenter la dimension logique des documents indépendamment de leur affichage [Michard, 98]. La figure suivante présente la dimension logique sous forme d'arbre de document multimédia de la figure 1.1.

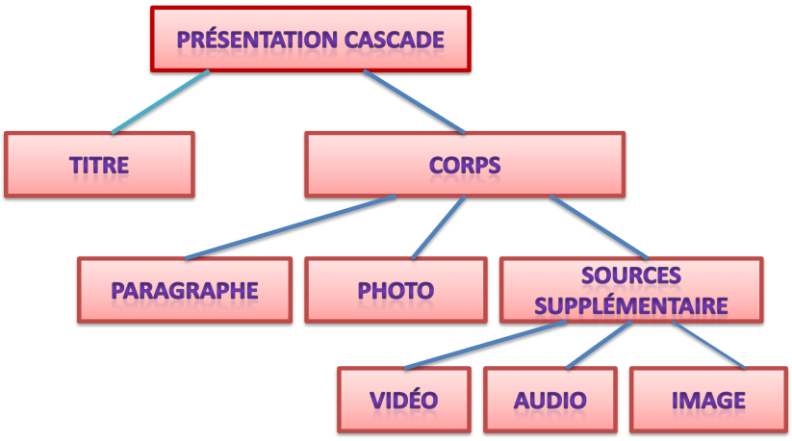


Figure 1.2. Dimension logique du document "Cascade thermale"

Le document est constitué d'un "titre" et d'un "corps", qui à son tour se compose d'un "paragraphe", d'une "photo" et des "sources supplémentaires". Cette dernière est composée d'une "présentation vidéo", d'un "enregistrement sonore", et d'une "illustration images".

4.1.2. Dimension physique

La dimension physique de document multimédia consiste à représenter les paramètres du signal propre à chaque type de média, par exemple, pour le média audio, on décrit la fréquence d'échantillonnage, la taille de l'échantillon, etc. [Jansen, 13]. Pour le média texte, sa dimension physique se définit par la mise en forme telles que les blocs de paragraphes, de pages, de couleur de texte, sa taille, etc. Le langage XSL (eXtensible Stylesheet Language) et les feuilles de style CSS (Cascading Style Sheets)[2] sont utilisés pour la mise en forme des pages web de format HTML. La dimension physique de notre exemple est illustrée dans la figure suivante.

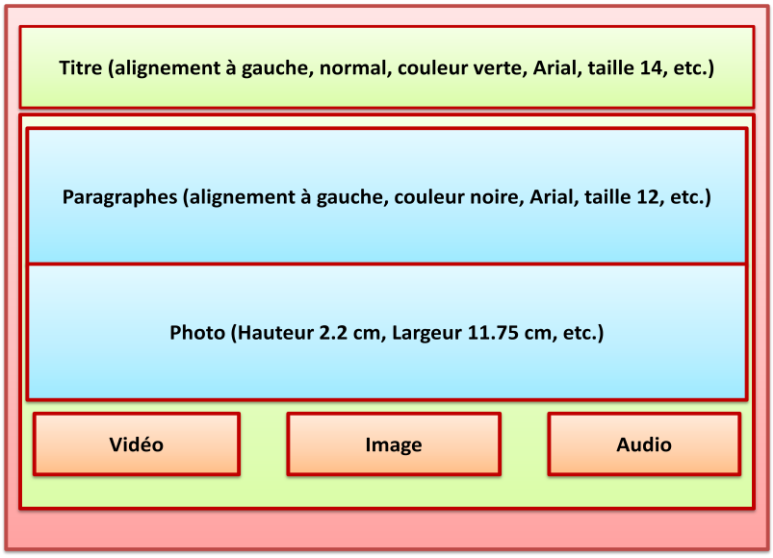


Figure 1.3. Dimension physique du document "Cascade thermale"

4.1.3. Dimension spatiale

La dimension spatiale permet de décrire le placement et les coordonnées des entités de document multimédia [Maredj, 13]. Elle permet donc de définir ces entités comme du texte, des vidéos ou des images. La dimension spatiale de notre document est la suivante: le titre occupe une largeur de 80% de celle du document et est centré ; les paragraphes sont situés après le titre, elles occupent 50% de l'espace total du document, l'image est placée juste après les paragraphes et les présentations vidéo, audio avec des illustrations images sont placées en bas de la page.

4.1.4. Dimension temporelle

La dimension temporelle est l'ensemble des informations qui décrivent la synchronisation des entités d'un document multimédia dans le temps et les relations temporelles qui existent entre elles [Jansen, 13]. Dans ce cadre, trois types de synchronisations apparaissent: la synchronisation intra-objets, la synchronisation inter-objets et la synchronisation des lèvres [Sabri, 13]. La synchronisation intra-objets s'applique aux relations temporelles qui existent entre les entités d'un média. La synchronisation inter-objets consiste à représenter l'enchaînement de la présentation de plusieurs entités.

[2] https://www.w3.org/TR/NOTE-XSL-and-CSS.html

La synchronisation des lèvres (lip-sync) est une combinaison des deux dernières. Elle impose un couplage temporel fort entre la progression temporelle de plusieurs entités.

Dans notre exemple de document multimédia, nous pouvons définir la structure temporelle comme suit :

1. L'apparition de la photo (la cascade) s'effectue en même temps que le titre du document et elle va durer jusqu'à la fin de l'affichage du document ou lorsqu'on choisit une source supplémentaire;
2. Supposons qu'on choisit les sources supplémentaires par ordre d'apparition. La présentation vidéo se lance après le clic sur son bouton. La vidéo va durer par exemple 15 minutes.
3. L'illustration images des artistes s'effectue 4 minutes après la présentation vidéo et elle va durer jusqu'au choix de la présentation audio.
4. La présentation audio se lance et elle va durer 10 minutes.

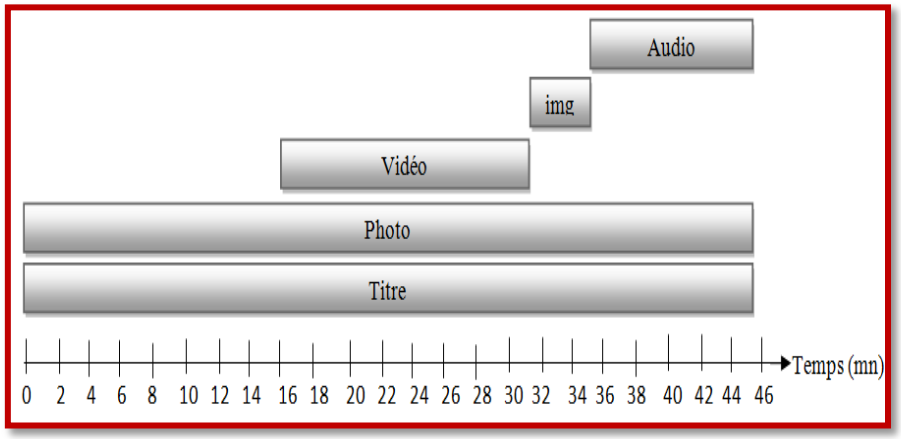


Figure 1.4. Dimension temporelle du document "Cascade thermale"

4.1.5. Dimension sémantique

La dimension sémantique permet de décrire le document multimédia par des entités de haut niveau montrant son contenu sémantique à l'aide d'un schéma [Hamadi, 15].

Le langage RDF (*Resource Description Framework*) est une recommandation du W3C développé pour décrire la sémantique des ressources du web à l'aide d'une structure à base d'XML [Lassila, 98].

Une structure sémantique pour le document "Cascade thermale" est présentée dans la figure suivante.

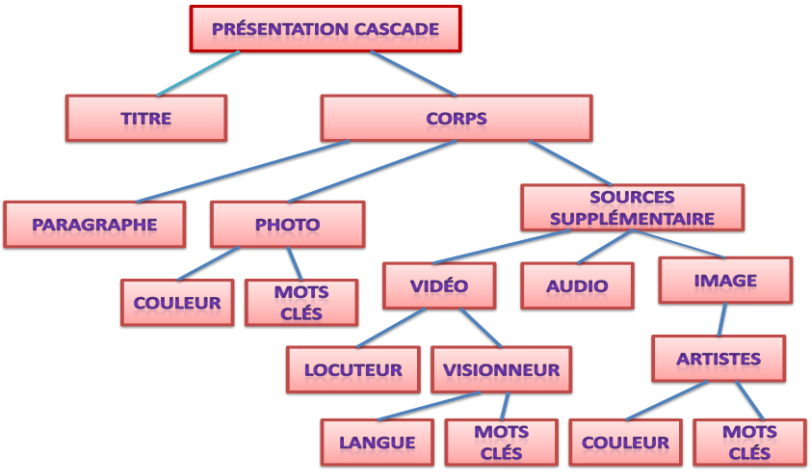


Figure 1.5. Structure sémantique du document "Cascade thermale"

Plusieurs normes et standard de l'organisation du contenu des documents multimédias, les plus populaires sont citées dans le travail de [Mbarki, 07]: SGML (Standard Generalized Markup Language) et XML (Extensible Mark-up Language) qui sont basés sur la représentation à base des balises, HyTime (Hypermedia/Time-based structuring language) qui prend en compte la synchronisation et la spécification d'hyperliens, SMIL (Synchronized Multimedia Integration Language) qui décrit l'organisation temporelle et spatiale des composants du document.

4.2. Base de données multimédia

Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités de données pour en faciliter l'exploitation (ajout, mise à jour, recherche de données) [Codd, 70] [Elmasri, 00]. Or, l'utilisation quotidienne de données multimédias et le besoin de les stocker, les organiser et les manipuler ont fait apparaître une nouvelle génération de bases de données, il s'agit les bases de données multimédias.

4.2.1. Définitions

Les bases de données multimédias (BDMM) s'appuient le plus souvent sur des architectures de base de données existantes [Gavini, 11]. Les plus utilisées étant le modèle relationnel et le modèle orienté objet. Les BDMM sont traitées initialement comme des bases de données stockant dans un attribut de chaîne de caractères le chemin vers les documents multimédias (documents textes, bandes sonores, fichiers images, vidéos) [Gavini, 11]. Elles ont par la suite la capacité de stocker les descripteurs de données multimédias à savoir les histogrammes de couleur pour décrire une base d'images [Sun, 14]. Les BDMM attirent l'attention des développeurs et des chercheurs dont le but d'assurer une représentation intégrale de ces données volumineuses au moyen des nouveaux types de données complexes utilisées dans la plupart des SGBD, appelés les types LOBs (Large Object) [Sun, 14].

Une base de données multimédia peut contenir des objets statiques (texte, images), des objets dynamiques où leur état évolue au fil du temps (son et vidéo) et des objets

dimensionnels (objets 3D). Selon le contexte d'utilisation de données, il existe deux types de bases de données multimédias [Sun, 14]:

- BDMM *génériques* dont le contenu est hétérogène (bases grand public, Internet, archives)
- BDMM *spécifiques* représentent une source de données d'un domaine d'application particulier telle que base de données botanique.

Une base de données multimédia doit assurer les fonctions suivantes [Gardarin, 03]:

1. Le stockage de tous les types de données multimédia (texte, audio, image et vidéo).
2. La manipulation de données multimédia via le langage de manipulation de données (LMD).
3. La gestion efficace de données volumineuses où sa taille peut atteindre les giga-octets.
4. La recherche par le contenu dont la requête peut contenir des objets multimédias.

À la lumière des définitions que nous venons de passer en revue, un système de gestion de bases de données multimédias (SGBD multimédia) doit supporter les quatre fonctions précitées en offrant un accès optimisé aux données. Le SGBD relationnel Oracle est un bon exemple pour définir et manipuler les BDMM. Dans cette optique, nous présentons dans la prochaine sous section les différents types de données permettant de présenter les données multimédias sous Oracle.

4.2.2. Types de données multimédias (Les LOBs)

À partir de la version 8 d'Oracle, des nouveaux types de données ont été intercalé pour stocker les données multimédias dans les colonnes de tables relationnelles [Sun, 14]. En outre, le langage d'interrogation SQL (Structured Query Language) est muni des nouvelles fonctionnalités pour supporter de manière intelligente des données multimédias. C'est à partir de la version 3 de SQL (baptisé SQL3) qu'on peut manipuler ce type de données [Eisenberg, 99]. Le SQL3 c'est une extension de SQL2 (développé en 1992) développée par le groupe de normalisation ANSI X3 H2 et internationalisée au niveau de l'ISO par le groupe ISO/IEC JTC1/SC21/WG3 [Eisenberg, 99]. Il est adopté en 1999 des nouveaux aspects, tels que l'intégration de l'objet au relationnel et l'intégration de multimédia via les types de larges objets, appelés les types LOB (Large Object) [Eisenberg, 99]. Selon la localisation des LOBs au niveau de la BDMM, nous pouvons distinguer deux types de LOB : LOB *interne* et LOB *externe* [Gardarin, 03] [Nwosu, 12]:

- *Les LOBs internes* sont stockés directement dans la table de la BDMM et référencés au moyen d'un pointeur logique appelé *Locator* qui est stocké dans la table de la base de données et qui pointe vers les données. Deux modes de représentation de LOB interne : la représentation binaire BLOB (Binary Large OBject) et la représentation en caractères longs CLOB (Character Large Object). Le BLOB est utilisé pour stocker les données binaires brutes de types vidéo, image et audio, dont la taille ne dépasse pas les quatre giga-octets. Le CLOB permet de stocker les données textuelles sous forme des chaînes

de caractères longs. Le NCLOB (National CLOB) est dédié pour les chaînes de caractères Unicode.

- *Les LOB externe* comme son nom l'indique, les données sont stockées dans un fichier externe et référencées à l'aide d'un pointeur. Le BFILE (Binary File) représente le type LOB externe où les données se trouvent dans un fichier externe à la base de données (sur disque dure, CDROM, etc.) et relient avec elle à travers l'attribut Locator qui pointe vers ce fichier.

Dans notre travail, nous nous focalisons sur le type BLOB et Image pour déclarer les données images et le type CLOB et Text pour la déclaration de données textuelles.

5. Stratégies d'exploration des sources de données multimédias

L'avènement de données multimédias et la disponibilité des outils de stockage représentent un nouveau défi pour la recherche ou l'exploration des sources de données multimédias. Pour rendre ces données facilement exploitables, il est nécessaire d'étudier les deux stratégies d'exploration des sources de données: recherche d'informations multimédias dans un corpus de documents et l'interrogation des bases de données multimédias.

5.1. Recherche d'informations dans un corpus de documents multimédias

La recherche d'informations dans un corpus de documents multimédias est une discipline à part entière de domaine de la recherche d'informations [Lazaridis, 13]. La recherche d'informations (RI) est un domaine lié aux sciences de l'information et à la bibliothéconomie, dont le but de retrouver des informations pertinentes dans un corpus pour un utilisateur ayant un besoin en information [Salton, 71]. De plus, un système de recherche d'informations (SRI) est « *un système d'informations qui permet de stocker l'information destinée à être traitée, recherchée, trouvée par une population variée d'utilisateurs* » [Salton, 89].

De ce fait, un modèle de recherche permet d'établir une représentation de la requête d'utilisateur, une représentation de documents et une fonction de correspondance entre les deux représentations [Moulin, 12]. Il existe trois principales familles de modèles de recherche: le modèle booléen [Salton, 71] qui a servi de point de départ aux recherches du domaine, il est basé sur la théorie des ensembles, dont la requête est composée de plusieurs termes reliés entre eux par les opérateurs de la logique booléenne (AND, OR, NOT), puis le modèle vectoriel [Salton, 71] c'est le modèle le plus souvent utilisé en RI, basé sur l'approche algébrique. Il est populaire grâce à sa capacité d'ordonner les documents retrouvés, sa robustesse et ses bonnes performances dans des tests. Enfin, le modèle probabiliste [Robertson, 76] [Maron, 60] fondé sur le calcul des probabilités, il permet de quantifier l'incertitude dans la représentation des informations ainsi que l'imprécision dans l'expression des besoins.

Quel que soit le type de données à rechercher dans un corpus de documents, le système de recherche d'informations se compose de deux principaux processus [Salton, 71]: le

processus d'indexation de documents et requêtes, et le processus d'appariement requêtes/documents.

5.1.1. Processus d'indexation

Le processus d'indexation consiste à représenter d'une manière unifiée à l'aide des index le contenu de documents et de requêtes afin de faciliter la comparaison entre la représentation d'un document et celle d'une requête [Maron, 60]. Ces index sont associés de poids obtenus généralement par la formule tf×idf, afin de représenter leurs degrés de représentativité du contenu sémantique de document (ou requête) qu'ils décrivent [Venturini, 14]. Quel que soit le mode d'indexation, manuelle (faite par un humain), semi-automatique (créer par un humain assisté d'un programme proposant des termes) ou automatique (créer par un programme informatique), l'indexation doit répondre à deux principaux problèmes, le choix des termes représentatifs de chaque document et l'évaluation de leur pouvoir de représentation[Baccini, 12].

Les données multimédias sont caractérisées par leur complexité et leurs paramètres temporels et spatiaux qui posent de nombreux problèmes d'indexation. L'indexation de données multimédias consiste à utiliser leurs descripteurs de bas niveau et de haut niveau [Feng, 13]. Pour l'indexation d'image, en utilisant des descripteurs de bas niveau tels que l'histogramme de la couleur, de la texture et la forme, ainsi que l'utilisation des ontologies pour la description de son contenu sémantique [Feng, 13]. Plusieurs projets d'indexation de données multimédias ont été développé, à titre d'exemple, le projet ANNAPURNA (ANNotation Automatique d'images PoUr la Recherche et la NAvigation) pour l'annotation des images [Chupeau, 03], le projet RAIVES (Recherche Automatique d'Informations Verbales Et Sonores) dédié à l'indexation des documents sonores [Parlangeau, 03], etc.

Concernant le processus d'indexation de documents textuels, quatre étapes ont été effectuées [Venturini, 14]:

1. *Analyse lexicale* : transformer un document textuel en un ensemble de termes (lexème). Pendant cette étape, la ponctuation, la casse, et la mise en page sont supprimées.
2. *Sélection* : consiste à garder que les termes discriminatifs pour un document, en utilisant par exemple un anti dictionnaire (tel que du système SMART) qui permet de ne pas conserver les mots qui ne reflètent pas le contenu informationnel des documents, à savoir : les articles, les pronoms, les prépositions, etc.
3. *Radicalisation :* consiste à éliminer les différences non significatives des mots (préfixes et suffixes) et de garder la partie commune appelée le radical (ou la racine).
4. *Pondération :* consiste à attribuer un poids à un terme d'indexation, de manière à représenter son pouvoir de discrimination pour chaque document de la collection.

5.1.2. Processus d'appariement

Afin d'extraire les informations satisfaisantes les besoins des utilisateurs, le processus d'appariement consiste à comparer les index de requête et ceux des documents par l'application d'une mesure de similarité entre la requête indexée et les descripteurs des documents du corpus [Salton, 71]. Seuls les documents dont la similarité dépasse un seuil prédéfini sont sélectionnés par le SRI. La fonction d'appariement ou de correspondance est un élément clé de SRI, car la qualité des résultats dépend de l'aptitude du système à restituer les documents pertinents les plus proches possibles au jugement de pertinence de l'utilisateur [Venturini, 14].

Il existe deux types d'appariement [Salton, 89]: appariement exact et appariement approché. Dans le premier type, le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés. Dans l'appariement approché du résultat est une liste de documents censés être pertinents pour la requête. Les documents restitués sont triés selon un ordre de mesure qui reflète le degré de pertinence document/requête.

Il fallait évaluer les SRI en fonction de leur capacité de retrouver l'ensemble de documents pertinents. Suite à une recherche, le corpus se divise en quatre ensembles de documents A, B, C et D [Tamine, 00].

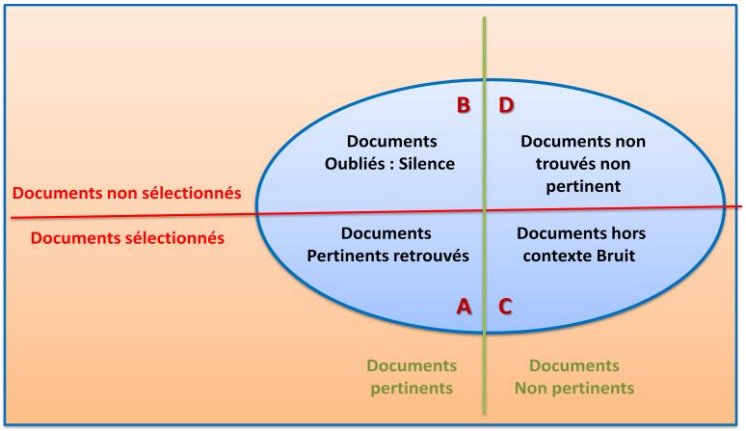


Figure 1.6. Les quatre ensembles de documents résultats en RI [Tamine, 00]

Le modèle d'évaluation utilisé en RI implique une collection de documents sur laquelle les recherches sont effectuées, un ensemble de requêtes de test et la liste des documents pertinents de la collection propre à chacune des requêtes [Salton, 89]. Ce modèle inclut également des mesures d'évaluation permettant de contrôler les performances des SRI [Baccini, 12]. Les mesures d'évaluation les plus utilisées sont : le rappel, la précision et la F-mesure. Nous nous basons sur les travaux de [Egghe, 08], [Baccini, 12] et [Tamine, 00] pour présenter ces mesures

- **Le rappel :** il mesure la proportion de documents pertinents retrouvés (l'ensemble A) parmi tous les documents pertinents disponibles dans la collection (l'ensemble A∪B). Si le rappel vaut 1 c'est que les documents pertinents disponibles ont tous été retrouvés par

le système, inversement si le rappel vaut 0 c'est qu'aucun document pertinent n'a été retrouvé. Cette mesure permet aussi de déterminer le silence, c'est-à-dire la proportion de documents pertinents non trouvés. Le rappel est donné par la formule suivante :

$$\textbf{Rappel} = \frac{Nombre\ \ de\ documents\ \ pertinents\ \ retrouv\ és}{Nombre\ \ de\ documents\ \ pert\ inent} = \frac{|A|}{|A|+|B|}$$

- **La précision :** elle mesure la proportion de documents pertinents retrouvés parmi tous les documents sélectionnés (l'ensemble A∪C). Elle mesure la capacité du système à trouver exclusivement des documents pertinents. La précision vaut 1 quand tous les documents retrouvés sont pertinents. Elle vaut 0 si aucun des documents retrouvés n'est pertinent. Cette mesure détermine également le bruit, c'est-à-dire la proportion de documents non pertinents restitués par le système. La précision se calcule alors de la manière suivante:

$$\textbf{Précision} = \frac{Nombre\ \ de\ documents\ \ pertinents\ \ retrouv\ és}{Nombre\ \ de\ documents\ \ sélectionn\ és} = \frac{|A|}{|A|+|C|}$$

En réalité, c'est rarement voir un système performant 100%, c'est-à-dire, il retourne pour une requête donnée seulement les documents jugés pertinents par l'utilisateur.

- **La mesure harmonique ou F-mesure :** L'idée essentielle de cette mesure est la possibilité d'avoir une valeur synthétisant le rappel et la précision. Ces deux mesures évoluent en sens inverse ; la précision est globalement décroissante au fur et à mesure que le SRI restitue des documents, alors que le rappel est globalement croissant. La F-mesure est calculée par l'expression suivante:

$$F = \frac{2 \times Rappel \times Précision}{(Rappel + Précision)}$$

5.1.3. Modes de recherche

La recherche de documents multimédias peut prendre quatre modes de recherche différenciés selon le type des couples (requête, documents): (texte, texte), (texte, multimédia), (multimédia, texte) et (multimédia, multimédia). Le premier mode de recherche est le mode le plus classique qui vise à rechercher un mot (ou un groupe de mots relient entre eux) dans un document [Smeaton, 12]. Par exemple, rechercher les livres de l'auteur 'Georges Gardarin'. Cette requête doit être indexée selon le processus d'indexation et par la suite l'application de la fonction de correspondance pour sélectionner les documents contenant le nom de cet auteur.

Le mode de recherche (texte, multimédia) permet de retourner des objets multimédias à partir d'une requête textuelle [Gao, 13]. Le résultat peut prendre plusieurs types, des images, des audio et des vidéos. La figure suivante montre l'exécution de la requête textuelle 'database' via le moteur de recherche Google[3] image. Le résultat de cette requête est un ensemble d'images signifiantes le mot database.

[3] https://www.google.com

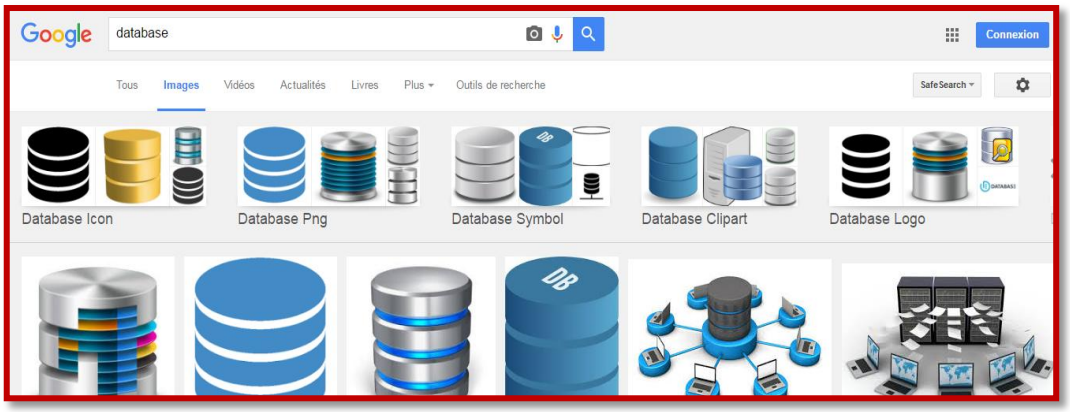


Figure 1.7. Mode de recherche (texte, multimédia) via Google image [1]

Le troisième mode de recherche (multimédia, texte) où la requête fournit des images exemples et utilisation des métadonnées textuelles associées à ces images pour retourner des données textuelles [Angus, 13]. On peut aussi y appliquer des requêtes vocales via un microphone.

Le quatrième mode de recherche est le mode purement multimédia ; il s'appelle aussi la recherche par le contenu. Ce mode de recherche est très connu en recherche d'images, qui est souvent appelé CBIR (Content Based Image Retrieval) [Gudivada, 95]. Le CBIR est une technique permettant de rechercher des images à partir de leurs caractéristiques visuelles, c'est-à-dire induite de leurs pixels [Gudivada, 95]. Les images sont classiquement décrites comme rendant compte de leur texture, couleur et forme. Un cas typique d'utilisation est la recherche par l'exemple où l'on souhaite retrouver des images visuellement similaires à un exemple donné en requête [Zloof, 77].

5.2. Interrogation des bases de données multimédias

Les bases de données multimédias permettent de représenter les données multimédias d'une façon structurée sous forme des tables relationnelles du modèle relationnel ou des classes d'objets issues du modèle orienté objet. Cette structuration assure un accès efficace aux données. De plus, le système de gestion de base de données multimédia (SGBD multimédia) est comme un SGBD classique représente par un ensemble de programmes puissants pour le stockage, la modélisation et l'interrogation des données [Mohammed, 14]. De plus, il doit supporter des fonctions avancées pour le stockage et le traitement de gros volumes de données multimédias.

L'intégration du multimédia dans les SGBD a été étudiée initialement par Woelk et Kim en 1987 [Woelk, 87], toutefois elle est véritablement débutée en 1992, incluse les problèmes de synchronisation et de traitement de documents hypermédias [Hoepner, 92] [Buchanan, 92].

Le SGBD multimédia offre donc les fonctions suivantes [Gardarin, 03] [Nwosu, 12]:

1. La description de données multimédias par l'utilisation de données complexe.
2. La recherche d'informations multimédias via une interface facilitant l'interrogation selon plusieurs modes de représentation (ex: requêtes, requête image).
3. La mise à jour de données multimédias.

4. Le contrôle de l'intégrité de données en spécifiant les valeurs permises pour certaines données, éventuellement en fonction d'autres données (ex: intégrité d'unicité de clé primaire, intégrité référentielle).
5. La gestion des transactions et sécurité.
6. La gestion des déclencheurs (Triggers) permet d'activer une procédure lors de l'apparition de conditions particulières dans la base de données.
7. La gestion des vues (Views).
8. L'optimisation des accès aux données en offrant des outils d'indexation spécialisés (Arbre B, R-tree).

L'architecture fonctionnelle de SGBD multimédia (cf. figure 1.7) comporte trois couches [Mohammed, 14] [Ahn, 12] [Djema, 07]: couche d'objets, couche de multimédia et couche de présentation.

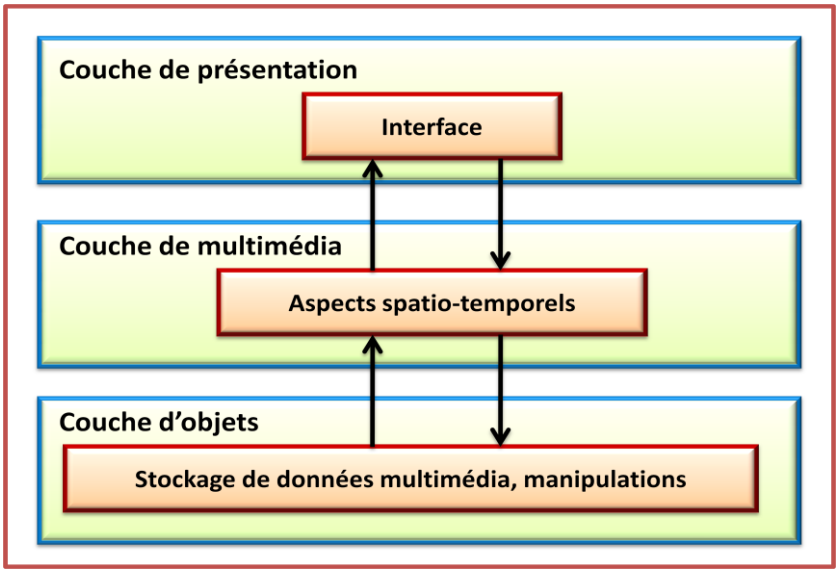


Figure 1.8. Architecture fonctionnelle d'un SGBD Multimédia [Djema, 07]

La couche d'objets est représentée par un ensemble de modules qui sert à décrire les aspects structurels des données multimédias et d'effectuer les manipulations usuelles (ajout, recherche, mise à jour et suppression). Elle offre aussi la possibilité de compresser les données multimédias afin de réduire l'espace de stockage et facilitant l'acquisition et la restitution de données. La couche de multimédia est la couche intermédiaire entre la couche d'objets et la couche de présentation. Elle est conçue pour traiter les aspects spatiaux et temporels de données multimédias. La couche de présentation permet de gérer les données multimédias d'une manière naturelle via une interface. Dans chacune des couches, on retrouve les trois niveaux : interne (ou physique), conceptuel et externe de données.

Le SGBD Multimédia est comme n'importe quel SGBD [Gardarin, 03], doté d'un langage de définition de données (LDD) permet de créer et de modifier l'organisation des données dans la base de données, langage de manipulation de données (LMD) permet de

rechercher, d'ajouter, de modifier ou de supprimer des données. Un langage de contrôle de données (LCD) permet d'autoriser ou d'interdire l'accès à certaines données aux personnes précises et un langage de contrôle de transaction (LCT) permet de commencer et de terminer des transactions.

5.2.1. Manipulation de types de données multimédias

Comme on a vu dans la section 4.2 que la déclaration de données multimédias dans une base de données est faite via les types LOBs. Nous présentons dans cette section les manipulations de LOBs via Oracle 10g que nous utilisons pour la création de nos bases de données relationnelles. Les LOBs interne (BLOB, CLOB et NCLOB) et les LOB externe (BFile) sont utilisés comme domaine d'une colonne de la table relationnelle, par exemple : image Blob, Vidéo Bfile. Les LOBs doivent être initialisés à vide afin de créer leur pointeur logique (Locator) au moyen d'une requête SQL [Menon, 05]. La création et la manipulation des BDMM oracle peuvent se faire de deux manières différentes [Menon, 05] [Lakshman, 03]:

1. *A travers le serveur oracle* : serveur de bases de données au moyen de SQL*Plus, ainsi le module d'oracle appelé Oracle intermedia qui fournit des classes pour stocker des données multimédias.
2. *A travers une application* : oracle fournit des bibliothèques dans quelques langages de programmation tel que java à travers le package oracle.sql.*.

Dans le cas de LOB interne, on prend comme exemple le type BLOB (la même chose avec CLOB). Pour manipuler les données BLOB, il est nécessaire de les initialiser à vide. L'initialisation du type BLOB est faite par la méthode *empty_blob()* et la récupération de son Locator grâce à la méthode *getBLOB()*[Allen, 09]. Une fois l'initialisation de type est terminée, on peut lire les données BLOBs. Dans le cas d'utilisation directe de serveur oracle, l'initialisation de LOB est effectuée à travers un programme PL/SQL (Procedural Language/SQL) [Allen, 09]. Dans le cas d'utilisation d'un programme Java, la BDMM doit être connectée au moyen de JDBC (Java DataBase Connectivity) sous forme de streams java et l'application de la méthode *getBinaryStream()* puis on lit la donnée grâce à la méthode *read()*[Lakshman, 03]. La lecture de Blob via oracle est réalisée par les deux expressions suivantes [Su, 12]: l'ouverture de blob par `dbms_lob.OPEN` et la lecture par `dbms_lob.READ`.

L'écriture de donnée BLOB est effectuée de la même façon que la lecture ; à la place de la méthode *getBinaryStream()* de java, on utilise la méthode *getBinaryOutputStream()* pour la récupérer sous forme de stream de sortie, puis on utilise la méthode *write()* à la place de *read ()*[Tarakanov, 15]. Dans un serveur oracle on utilise la méthode `utl_file.put` dans un programme PL/SQL [Tarakanov, 15]. De même, la suppression d'un BLOB supprime à la fois les données BLOBs et leur pointeur logique [Tarakanov, 15].

Exemple : nous voulons créer une base de données multimédia sous le SGBD Oracle. Cette base comporte une seule table relationnelle nommée T_Food qui contient deux attributs, la clé primaire Num est un numéro séquentiel et l'attribut Food représente l'image de Food dont son type est Blob. Avant de créer cette table, il faut assurer la connexion avec le serveur d'Oracle

puis on crée un répertoire (DIRECTORY) nommée foodDIR qui nous servira à placer les images de Food afin de pouvoir accéder à ces images.

```
SQL> CREATE OR REPALCE directory foodDIR AS 'd:/imageF/';
Directory created.
SQL> GRANT READ ON directory foodDIR TO public;
GRANT succeeded.
```

L'expression suivante permet de créer la table T_Food:

```
CREATE TABLE T_Food (Num number not null primary key, Food blob);
```

Pour insérer l'image Fish.jpeg dans cette table, il fallait de créer une procédure d'insertion PL/SQL, nommée par exemple add-food puis on l'exécute grâce à l'expression :

```
exec sql_blob.add_food(1,'Fish.jpeg');
```

Le contenu de procédure add-food prend comme paramètres les mêmes types d'attributs de la table T_Food.

```
PROCEDURE add_food (id NUMBER, name BLOB) IS
    v_food BLOB;
    v_file BFILE;
  BEGIN
    -- Initialisation de type blob
    INSERT INTO T_food
    VALUES
      (id, empty_blob())
    RETURNING Food INTO v_food;
    -- déclaration de pointeur vers le fichier v_file
    v_file:= bfilename('foodDIR', name);
    -- On ouvre ce fichier
    dbms_lob.fileopen(v_file);
    -- remplissage de variable de type BLOB par le contenu de fichier
v_file
    dbms_lob.loadfromfile(v_food, v_file, dbms_lob.getlength(v_file));
    -- fermeture de fichier
    dbms_lob.fileclose(v_file);
  END;/
```

En ce qui concerne le LOB externe, le type BFILE utilise le pointeur logique *Locator* qui pointe vers un fichier externe en dehors de la base de données ainsi que la suppression d'un BFILE supprime seulement le pointeur, mais pas le fichier qui était référencé [Tarakanov, 15]. De plus, les données de BFILE sont read-only et ne peuvent pas insérer de données ni écrire dans un BFILE, l'administrateur de la base de données doit s'assurer que la lecture est la seule opération permise sur le fichier [Tarakanov, 15]. L'intégrité des données n'est plus assurée par Oracle mais par le système d'exploitation. La déclaration de type BFILE est comme n'importe quel type de donnée, il suit le nom de l'attribut [Menon, 05]. Prenons l'exemple précédent de la table T_Food, l'attribut Food est du type BFILE :

```
CREATE TABLE T_Food (Num number not null primary key, Food bfile).
```

La manipulation des types BFILE est la même que les LOB interne, elle s'effectue soient via le serveur oracle ou le code java, ainsi, la création d'un BFILE est beaucoup plus simple que celle des LOB interne, elle s'effectue simplement au moyen de la méthode bfilename ('nom_répertoire' , 'nom_donnée') [Menon, 05]. Par exemple, pour insérer une nouvelle ligne avec un pointeur vers le fichier Meat.jpeg, on exécute la requête d'insertion suivante :

```
INSERT INTO T_food VALUES (1, bfilename('foodDIR', 'Meat.jpeg')); COMMIT;
```

Par ailleurs, le SGBD Oracle fournit une suite des services appelée Oracle intermedia depuis sa version 8 (en 1997), pour gérer les bases de données multimédias [Allen, 09]. Oracle intermedia est constitué d'un package ORDSYS ("ORD" pour les données objet-relationnelles) permettant la gestion des objets multimédias dans la base. Il comprend plusieurs classes [Su, 12]:

- *ORDMultimedia* : superclasse abstraite stockant les attributs et méthodes communs aux classes ORDAudio, ORDImage, et ORDVideo.
- *ORDAudio* : permet de stocker des médias du type audio et disposer des méthodes pour gérer et rechercher des objets sur différents critères (métadonnées).
- *ORDImage* : supporte l'ensemble de méthodes (compression, rotation, contraste, symétrie, segmentation) et des attributs propre à une image (couleur, forme, texture, localisation, poids, score, seuil).
- *ORDVideo* : permet de traiter, de stocker, de diffuser (streaming) et de calculer les propriétés d'un fichier vidéo.
- *ORDDoc* : principalement utilisé pour les documents Word et HTML. Il contient les attributs et les méthodes de base afin de manipuler légèrement une image, un audio ou un extrait vidéo.
- *ORDSource* : stockage des sources multimédias dans des BLOB de la base, ou des BFILE.

Oracle interMediaText est un module Oracle dédié aux traitements de données textuelles [Tarakanov, 15]. Il offre des facilités pour gérer et rechercher sur le texte. LONG est le type d'attribut textuel qui permet de stocker jusqu'à 2 Go de caractères [Tarakanov, 15]. De plus, lorsque la colonne de texte est indexée les fonctionnalités de recherche sont enrichies.

5.2.2. Modes d'interrogation de BDMM

Les bases de données multimédias offrent la possibilité de stocker et d'interroger des données multimédias. De même, les langages d'interrogation comme SQL doivent être étendus afin de gérer ce type de donnée. En 1992, une réunion des comités de normalisation SQL a eu lieu à Tokyo, dont le but d'étendre le SQL par l'ajout des extensions orientées objet à SQL ce qui est souvent appelé le *SQL3* [Eisenberg, 99]. Ce dernier a été mise en route en 1999 et le modèle relationnel devient "*modèle objet-relationnel*" où un n-uplet de base de données peut être une valeur ou un objet ayant un OID (Object Identifier) permettant de référencer une table objet-relationnelle [Melton, 01]. Le SQL3 intègre les types de données définissables par l'utilisateur qui sont appelés types abstraits de données ***ADT*** (Abstract Data Type) et qui sont représentés par la syntaxe suivante [Mattos, 99]:

```
CREATE TYPE <nom ADT> AS <corps de l'ADT>;/
```

Le langage SQL3 est comme n'importe quel langage d'interrogation, composé d'un certain nombre des parties (ou composants) modulaires ; les six premières sont de base [Mattos, 99]: Framework (description de la structure du document), Foundation (Le noyau de spécification, incluant les ADT), SQL/CLI (l'interface d'appel client), SQL/PSM (le langage de spécifications de procédures stockées), SQL/Bindings (les liens SQL dynamique et SQL repris de SQL-92) et le SQL/XA (Une spécification de l'interface XA pour moniteur transactionnel). Une partie de SQL3 est destinée aux données multimédias, s'appelle **SQL/MM** (SQL/MultiMedia) qui y comprit principalement les données en texte intégral (Full Text), les données spatiales (Spatial standard) et des données d'image (Still image) [Mattos, 99].

En général SQL/MM est fondée sur les types LOB pour définir les données multimédias [Mattos, 99]. L'interrogation de ce type de donnée peut être réalisée selon deux modes d'interrogations: interrogation textuelle et interrogation visuelle. Nous nous basons sur les travaux de [Mattos, 99], [Melton, 01] et [Yen, 93] pour présenter ces modes d'interrogation.

Interrogation textuelle : il s'agit d'interrogation classique exploitant un langage de requêtes tel que SQL. De ce fait, la spécification des éléments d'une requête dépend de la connaissance des données et de leurs structurations (nom des tables, attributs). Dans le cas de définition de données textuelles, on peut utilisées par exemple le type FULLTEXT. Ce dernier supporte des méthodes sur ses données et il permet de les convertir en type ordinaire de SQL sous forme des chaînes de caractères. Les index de données FULLTEXT pourraient enregistrer des informations sur la proximité des mots et des phrases les unes aux autres ou sur les mots qui apparaissent dans un document et des mots liés qui ne figurent pas dans le même document. Le type FULLTEXT comporte plusieurs méthodes à savoir, concaténation de deux mots, longueur d'une expression, position d'un mot dans un texte, recherche d'un mot dans un texte, conversion en chaîne de caractères, etc.

Exemple : soit la table objet-relationnelle suivante :

```
CREATE TABLE T_Livre (Num number not null primary key, titre VARCHAR2(40),
texte FULLTEXT);
```

La requête suivante permet d'afficher le titre des livres qui possèdent le mot 'data' dans le même paragraphe contenant les mots qui se prononcent pareils au mot «base».

```
SELECT titre
FROM T_Livre
WHERE texte.CONTAINS
      ('STEMMED FORM OF "Data"
       IN SAME PARAGRAPH AS
       SOUNDS LIKE "base"') = 1
```

La méthode CONTAINS est une fonction booléenne, vaut 1 qui signifie que le résultat doit être vrai pour les conditions passées en paramètres. Ces derniers sont liés entre eux à travers trois opérateurs suivants : `STEMMED FORM OF "Data"` permet de trouver tous les mots dérivés au mot Data tels que database, dataset, metadata. L'opérateur `IN SAME PARAGRAPH AS`, exige qu'un deuxième mot (ou une phrase) apparait dans le même paragraphe que le mot data. L'opérateur `SOUNDS LIKE` trouve les mots qui sont prononcés (en anglais) à un mot donné.

Le SQL/MM Spatial standard [Stolze, 03] permet de représenter les données spatiales et géométriques comme les points, cercle, carré, polygone grâce aux types ST_Geometry qui se divisent en deux types : les types simples (ST_ point, ST_ curve, ST_ surface) et les types complexes (ST_Multipoint, ST_Multicurve, ST_Polygon, ST_Path, …etc). Les types ST_Angle et ST_Direction, sont employés pour stocker les informations sur les angles et diverses directions qui sont nécessaires quand en stocke et en gère les données spatiales. Le SQL/MM offre plusieurs opérations pour interroger les données spatiales via une requête SQL [Stolze, 03], à titre d'exemple, la construction d'une ligne droite à partir de deux points, construction du polygone à travers plusieurs lignes et plusieurs points, ou à partir d'un ensemble de directions et distances, etc. La requête suivante permet d'afficher la localisation spatiale (par la méthode area) de faculté des sciences de la table relationnelle université :

```
SELECT Emplacement.area
FROM   T_université
WHERE  Faculté ='Sciences';
```

Avec `Emplacement, Faculté` sont des attributs du type ST_Geometry et VARCHAR2(30) respectivement.

Ce type de donnée est très utilisé dans le système d'informations géographiques (GIS). En ce qui concerne la représentation des images (fixes et animées), le package SQL/MM Still image comporte un ensemble de types de données images qui sont appelés SI_StillImage [Dessloch, 97]. Ces types permettent de stocker les caractéristiques visuelles de l'image telles que : couleur, texture, forme, format, dimension,…etc. Nous pouvons citer quelques types SI_StillImage : SI_AverageColor, SI_ColorHistogram, SI_PositionalColor, SI_Texture, SI_FeatureList.

Par ailleurs, oracle intermedia permet d'interroger les données ORDSYS via une requête SQL, par exemple pour insérer une image ORDSYS.ORDImage, il est nécessaire de

l'initialiser via la méthode init (Type de source, répertoire, nom de l'image) [Su, 12]. Dans la table T_food présentée dans la sous-section précédente ; l'insertion de l'image Meat.jpeg qui se trouve dans un fichier externe du répertoire `foodDIR` est faite comme suit:

```
INSERT INTO T_food VALUES (1, ORDSYS.ORDImage.init('file','foodDIR',
'Meat.gif')5
```

En 2015, Lu et al [Lu, 15] ont proposé une extension du langage SQL dédiée à l'interrogation des bases de données des vidéos, ce langage s'appel SVQL (Structured Video Query Language) qui apporte des nouvelles conditions dans la clause WHERE telles que la déclaration des variables, la déclaration des entités vidéo, spécification de structure, spécification des caractéristiques et la spécification spatiotemporelle.

D'une façon générale, l'interrogation de données multimédias exclue certaines opérations de base telles que > et<, ainsi qu'on ne peut pas définir un objet multimédia comme clé primaire ou étrangère [Su, 12].

Interrogation visuelle : elle consiste à utiliser des objets graphiques comme requête présentée au niveau d'une interface graphique [Zloof, 77]. Cette interrogation est faite en donnant un "motif" qui peut être complété par des informations correspondantes. Ce motif peut être une image, un son, une vidéo. L'interrogation par l'exemple baptisée QBE (Query By Example) permet à un utilisateur quelconque de rechercher des données dans une base de données en donnant la structure de la table résultante, à travers d'une interface graphique [Yen, 93]. QBE s'appuie aussi sur le calcul relationnel tout en offrant aux utilisateurs la convivialité de l'approche graphique pour travailler avec les tables. Il a été inventé par Moshe Zloof pour le compte de la compagnie IBM, en 1977 [Zloof, 77]. Il a connu un succès grâce à son introduction au sein de la première version de Paradox (1.0 pour DOS) en 1985 [Yen, 93]. Le QBE est un langage non procédural, doté d'une interface graphique, avec lequel l'utilisateur conçoit et exécute ses requêtes [Zloof, 77]. Au début, le QBE est fondé sur la visualisation de contenu de la base de données par la commande DRAW nom-de-table. Le résultat de cette commande est une représentation graphique de la table relationnelle avec tous ses attributs [Aversano, 02]. Par exemple, l'exécution de la commande DRAW T_FILM nous a permis d'afficher la table T_FILM suivante :

T_FILM	N#	Acteur	Année	Titre du film	Image
P.			2016	N	

Table 1.1. Exemple d'une requête graphique sous QBE

La table ci-dessus représente une requête graphique qui permet d'afficher tous les films produits en 2016, dont leur titre commence par la lettre N. La commande P. (signifie *Print*) permet de sélectionner les attributs que nous voulons les afficher. Dans cet exemple, la commande P. est mise sous le nom de la table T_FILM afin de sélectionner tous ses attributs. Pour définir plusieurs conditions sur le même attribut, nous utilisons l'opérateur AND ou OR (remplacer par une nouvelle ligne dans la table de requête). Pour insérer un nouveau tuple, on met au-dessous du nom de la table la commande I. (signifie *Insert*) et par la suite on remplit

les colonnes avec les valeurs des attributs. On procède la même stratégie pour effectuer la mise à jour via la commande U. (signifie *Update*) et la suppression par la commande D. (signifie *Delete*).

Dans le cas des bases de données multimédias, la requête est elle-même une donnée multimédia, à savoir une image requête et ses résultats correspondent à une liste d'images ordonnées en fonction de la similarité [Aversano, 02]. L'interrogation visuelle est très utilisée dans le domaine de l'imagerie ; à partir d'une image qui a été sélectionné comme requête, son exécution vise à mesurer la similarité entre les images de la BDMM et une image requête [Yen, 93]. Autrement dit, plus la distance entre les deux est petite, plus la similarité est grande, donc plus elles sont similaires. Il existe plusieurs mesures de similarité, à titre d'exemple la similarité cosinus, Wu and palmer, etc. Le choix d'une telle mesure dépend beaucoup de la manière avec laquelle l'image est recherchée. Par ailleurs, l'interrogation visuelle permet aussi d'utiliser une esquisse (dessin) comme requête d'utilisateur [Yen, 93]. L'esquisse peut être une ébauche de forme, une texture ou une ébauche de couleur [Braga, 03]. Néanmoins, cette technique présente l'inconvénient majeur pour définir une esquisse. La mise en œuvre de QBE dans la pratique a montré que la lisibilité des requêtes complexes formulées dans ce langage est moins bonne que celle des instructions SQL [Braga, 03]. D'autres langage d'interrogation visuelle, concerne l'utilisation des images picturales comme requête à l'aide de langage QPE (Query by Pictorial Example) qui est très utilisé pour l'extraction d'un croquis (ou schémas) à partir d'une image satellite [Chang, 80].

Ces deux modes d'interrogation doivent être prendre en compte des métadonnées caractérisant le contenu sémantique de données multimédias. Dans ce cadre, la version 12c d'Oracle offre un outil de gestion des thésaurus via Oracle Text thesaurus afin d'optimiser les recherches dans les bases de données sur le web [Das, 15]. Le thésaurus est un ensemble de termes bien déterminés pour un domaine donné, relient entre eux par des relations sémantiques et génériques [Moreira, 04]. Les relations sémantiques entre les termes du thésaurus peuvent être utilisées dans une condition CONTAINS. Par exemple, CONTAINS (*attribut*, BT (*mot*, *niveau*, *thesaurus*) > 0. Avec *attribut* est la colonne de recherche, *mot* est le mot cherché, *niveau* indique la profondeur de recherche dans l'arbre, *thesaurus* est le nom du thesaurus. Grâce aux outils d'oracle, l'utilisateur libère de la nécessité de saisir manuellement les commandes SQL. Les sources de données multimédias sont très divers tant par leur structure que par leur sémantique et leurs types de données. L'accès aux sources de données multimédias et hétérogènes devient un problème d'actualité qu'on devra l'étudier et de proposer une solution efficace.

6. Problèmes liés aux données multimédias

À la lumière des définitions que nous venons de présenter, les données multimédias sont par nature volumineuses et hétérogènes et leur contenu est généralement visuel qui ne permet pas de représenter aisément leur contenu sémantique. De plus, ces données sont représentées et stockées dans une multitude de sources de données qui sont le plus souvent autonomes, distribuées et hétérogènes. Par conséquent, des problèmes liés aux données multimédias sont à la base de nombreux travaux de recherche, les plus connus sont :

- ***Problème du stockage*** : les données multimédias sont volumineuses et nécessitent des outils puissants pour leur stockage. La plupart des solutions existantes visent à compresser les données multimédia afin de diminuer l'espace de stockage requis et de permettre l'exploration efficace de données [Sharma, 15]. D'autres travaux consistent à utiliser les index qui permettent d'optimiser le traitement de requêtes dans une BDMM en réduisant les accès disques [Shah, 14]. Les principales techniques d'indexation sont basées sur les R-Tree qui sont des extensions des index B-arbres [Shah, 14]. Ainsi, que les sources de stockages sont généralement hétérogènes. Elles peuvent être des BD relationnelles, BD orientée-objet, corpus de documents, site web, etc.
- ***Problème de modélisation*** : la modélisation de données multimédias est un travail subjectif, dans le fait que les mêmes données peuvent être définies par différents concepteurs avec différents schémas (XML, modèle relationnel, modèle orienté-objet) et diverses structures. De plus, des expressions distinctes ont été utilisées pour représenter les mêmes données et des logiciels différents sont utilisés pour créer et gérer les données (e.g. Oracle, O2, MySql). Des travaux de modélisation de contenu sémantique de données multimédias via des métadonnées ont été réalisés et qui peuvent être classifiés en deux familles : i) les travaux de modélisation séparée des médias : leur principe est de proposer un modèle particulier pour chaque type de média en utilisant une liste prédéfinie des descripteurs [Amous, 02]. ii) la seconde famille comprend des travaux, plutôt, basés sur une modélisation globale des médias: leur principe est de proposer un modèle détaillé et générique des différents éléments et de leurs relations [Jedidi, 05].
- ***Problème d'hétérogénéité sémantique*** : L'hétérogénéité sémantique est un problème très connu dans le contexte de données textuelles où les mots composant la requête ne permettent pas de définir correctement les besoins des utilisateurs. Ce problème s'augmente quand les données sont de natures multimédias, à titre d'exemple deux images ayant les mêmes caractéristiques visuelles mais elles sont sémantiquement différentes. Dans le domaine du web sémantique, une solution consiste à associer des ressources sémantiques les plus souvent sont les ontologies avec les sources de données hétérogènes [Jain, 13]. L'intégration de la sémantique en recherche d'informations (RI) donne lieu une nouvelle discipline de la RI appelée la recherche sémantique d'information (RSI) [Jain, 13]. De même, une solution de bases de données à base ontologique (BDBO) répond au mieux aux besoins et permet de traiter ce problème d'hétérogénéité [Mbaiossoum, 12]. Les éléments du schéma de BDBO sont alors liés à une ontologie pour en définir la sémantique.
- ***Problème d'exploration de données*** : plusieurs techniques de recherche ou d'exploration de données multimédias, les plus connues sont la recherche dans un corpus de documents suivant les processus du système de recherche d'informations et l'interrogation des bases de données multimédias BDMM. Dans les deux stratégies de recherche, on peut y avoir une exploration textuelle ou visuelle en appliquant une mesure de similarité entre la requête multimédia (par exemple image requête) et la collection de données multimédias. La recherche et l'interrogation de données issues des sources hétérogènes nécessitent la mise en place d'un modèle unifié assurant un

accès efficace aux différentes sources de données. Dans ce contexte, deux approches ont été proposées dans la littérature pour assurer un accès unifié aux sources de données [Vidal, 13] [Khouri, 12] : une approche de médiateur (ou intégration virtuelle) qui préconise la définition d'un schéma global virtuel sur lequel on soumet les requêtes d'utilisateur et l'approche d'entrepôt de données (intégration matérialisée) qui préconise la centralisation de toutes les données des sources au niveau d'une localisation unique.

Dans le cadre de notre travail, nous avons mis l'accent sur ces différents problèmes liés aux données multimédias et plus précisément le problème d'hétérogénéité sémantique pour assurer un accès unifié aux sources de données multimédias. Afin d'accéder aux données pertinentes issues des sources de données multimédias et hétérogènes, nous avons alors besoin de les traiter.

7. Conclusion

L'objectif de ce chapitre était de présenter les données multimédias, leurs descripteurs (descripteurs bas niveaux et descripteur de haut niveau) et leurs représentations (documents multimédias et bases de données multimédias). Afin de rendre ces données facilement exploitables, nous avons distingué deux stratégies de recherche ou d'exploration de données: recherche d'informations dans un corpus de documents multimédias et interrogation des bases de données multimédias. Les données multimédias sont souvent issues des sources de données hétérogènes ce qui peuvent générer des problèmes influent sur la qualité des résultats retournés.

Les données sur lesquelles nous concentrons notre étude sont issues des bases de données hétérogènes contenant à la fois des données ordinaires, des textes et des images. De ce fait, l'approche d'intégration que nous prenons en compte dans notre étude est l'approche de médiation et plus particulièrement l'approche de médiation sémantique à base d'ontologie. D'un autre côté, nous utilisons les documents multimédias scientifiques intégrant des textes et des images pour la proposition d'un nouveau modèle d'indexation de documents.

Dans le chapitre suivant, nous présenterons les différentes approches d'intégration des sources de données hétérogènes et en mettre l'accent sur les approches de médiation sémantiques à base d'ontologie.

CHAPITRE

02

HÉTÉROGÉNÉITÉ SÉMANTIQUE ET INTÉGRATION À BASE DE MÉDIATEUR

Ce chapitre porte sur les approches existantes pour le traitement du problème d'hétérogénéité sémantique des sources de données. L'idée principale de ces approches est d'offrir un système d'intégration de données permettant d'assurer à l'utilisateur un accès unifié aux données sans besoin de connaître leurs sources d'origine. Ce chapitre présente dans un premier lieu les problèmes liés à l'hétérogénéité sémantique et dans un deuxième lieu les approches d'intégration de sources de données hétérogènes. Il présente ainsi, de manière plus générale, les approches de médiation sémantique à base d'ontologie et de manière plus particulière, la médiation sémantique des sources de données multimédias.

1. Introduction

Le besoin d'exploration des sources de données est de plus en plus fort. Ces sources sont souvent hétérogènes, contenant des données multimédias. Un des verrous scientifiques à lever concerne le traitement des problèmes d'hétérogénéité sémantique des sources de données multimédias et hétérogènes. De ce fait, l'intégration sémantique de sources de données consiste à intégrer ces sources dans un contexte d'enrichissement sémantique.

Notre travail porte sur le traitement d'hétérogénéité sémantique selon deux principaux axes : (1) l'exploration des bases de données multimédias en fournissant une intégration sémantique, et (2) l'exploration de documents multimédias en proposant une nouvelle approche d'indexation de ce type de document. Pour présenter notre travail selon le premier axe, ce chapitre présente un état de l'art sur les travaux relatifs à l'intégration des sources de données et plus précisément les données multimédias.

Dans ce chapitre nous allons présenter les problèmes liés à l'hétérogénéité sémantique qui sont dues à la diversité des bases de données multimédias. Nous décrivons les différentes approches d'intégration de données, d'une part, et de mapping entre le schéma du système de médiation et les schémas locaux des sources à intégrer d'autre part. Nous nous sommes focalisés sur l'approche d'intégration via le système de médiation et nous détaillons son architecture de trois couches (médiateur, adaptateurs et sources de données). Nous situons également la place des ontologies dans le processus d'intégration pour le traitement du problème d'hétérogénéité sémantique des sources de données et nous discutons les différentes approches de médiation sémantique à base d'ontologie. Enfin, nous visons à présenter les travaux réalisés dans le cadre d'intégration des sources de données multimédias.

2. Notions fondamentales

Dans cette thèse, nous étudions les problèmes d'hétérogénéité sémantique au sein des données issues des sources hétérogènes. Nous nous sommes focalisés à traiter ces problèmes dans le cas de données multimédias en assurant une meilleure intégration de ce type de données. Après avoir présenté dans le chapitre précédent, les concepts de base liés aux données multimédias, nous présentons dans cette section d'une part, les problèmes d'hétérogénéité sémantique liés à la diversité des sources de données et d'autre part l'intégration de ces sources.

2.1. Problèmes liées à l'hétérogénéité sémantique

Avec l'essor technologique, les données multimédias sont issues de plusieurs bases de données (BD) qui sont généralement hétérogènes, que ce soit en termes des SGBD utilisés (Oracle, O2, MySQL, Access, etc.), les schémas de modélisation (relationnel, orienté-objet, déductif,…) ou les formats de résultat (objets, n-uplets, image, etc.). Cette diversité des sources de données peut conduire à une situation d'hétérogénéité sémantique dans le cas où la représentation de données est un travail subjectif qui dépend des pensées distinctes des concepteurs des BD.

Ainsi, plus le système est hétérogène, plus il est difficile d'assurer l'interopérabilité en son sein. Dans le domaine des bases de données fédérées, l'hétérogénéité sémantique se définit par « *les différentes significations, interprétations et utilisations souhaitées des données semblables ou liées* » [Sheth, 92].

L'intégration des bases de données multimédias donne lieu à de nombreux problèmes liés à l'hétérogénéité sémantique, les plus souvent sont [Aggoune, 15a] [Aggoune, 17] :

- ***Problème d'unification du modèle de données:*** avec la richesse des modèles de représentation de données à savoir, le modèle relationnel, orienté-objet, logique, etc. Les concepteurs ont le choix d'utiliser un tel modèle pour représenter leurs données. Ce problème a conduit à d'autres problèmes liés aux différents langages de manipulations de données ainsi que le même langage comme SQL peut être utilisé dans divers SGBD (Microsoft Access, Oracle, MySQL, InformixSQL, PostegreSQL, etc.). Il fallait offrir un modèle unifié pour pallier à la fois aux problèmes d'intégration des différents modèles et d'utilisation des divers systèmes de gestion des bases de données (SGBD).
- ***Problème d'unification de structure de données:*** avec le même modèle de données à titre d'exemple le relationnel, on peut avoir plusieurs structures (ou schémas) différentes, puisque les concepteurs des BD ne possèdent pas les mêmes pensées et n'ont pas forcément les mêmes idées pour la structuration de données. En effet, les structures de données se diffèrent entre elles selon le nombre des relations, des attributs et leurs types, le degré de décomposition des attributs et les contraintes faites sur les données. Il devra comparer les composants des schémas de données en utilisant une mesure de similarité entre eux.
- ***Problème de standardisation de représentation:*** quel que soit le type de données multimédias manipulées (texte, image, audio et vidéo), il est indispensable d'associer des termes ou même de les utiliser comme nom des entités (relations, attributs, méthodes et contraintes). L'hétérogénéité sémantique des données multimédias porte sur les différentes annotations ou descriptions pour le même objet multimédia. Il porte aussi sur l'utilisation des mêmes données ordinaires (chaîne de caractères, entier, …) avec différentes données multimédias dans des bases de données multimédias. De plus, la diversité d'expressions est due aux deux aspects de la langue naturelle (cf. chapitre 1, Section 2.1): la synonymie et la polysémie, qui peuvent provoquer des problèmes d'ambiguïté des mots. De plus, le même terme peut être défini par son nom d'origine ou par son abréviation qui peut générer des problèmes d'incohérence et de redondance. On peut y avoir des termes syntaxiquement incorrects ou incomplets qui ne permettent pas de spécifier correctement le sens du mot.
- ***Problème d'accès aux données:*** l'interrogation des bases de données est devenue une tâche omniprésente pour les utilisateurs. Toutefois, cette tâche devient de plus en plus complexe à cause de la diversité des langages d'interrogation propres à chaque base de données (SQL, OQL, QBE, etc.). En effet, les utilisateurs sont confrontés au problème d'expression de leurs besoins qui met à jour les difficultés de retrouver les réponses pertinentes désirées par l'utilisateur. Ce problème s'augmente quand la requête comporte des objets multimédias. Il faut donc, mettre en place un langage standard pour

accéder facilement aux sources de données tout en assurant la transparence à l'hétérogénéité.

En grosso modo, tous ces problèmes d'hétérogénéité sémantique restent toujours existants et il est impossible d'avoir des sources de données homogènes, conçues par différents concepteurs et adaptées aux besoins de tous les utilisateurs. Néanmoins, pour faire face à ces problèmes, des systèmes d'intégration de données hétérogènes ont vu comme un standard permettant d'assurer à l'utilisateur un accès unifié aux données sans besoin de connaître leurs sources d'origine [Lenzerini, 02]. En effet, il est nécessaire de découvrir la définition du concept d'intégration de données et les évolutions des systèmes dédiés à l'intégration de sources hétérogènes.

2.2. Intégration des sources de données hétérogènes

L'intégration des sources de données hétérogènes ou tout simplement *l'intégration de données* provenant de sources hétérogènes est un besoin crucial dont le but de donner à l'utilisateur l'illusion de n'interagir qu'avec des sources de données homogènes [Wiederhold, 92]. Lenzerini Maurizio [Lenzerini, 02] définit le concept d'intégration de données comme suit: « *L'intégration de données est le problème qui consiste à combiner des données qui résident dans différentes sources, et de fournir à l'utilisateur une vue unifiée de ces données* ». Le système d'intégration des sources de données hétérogènes (SID) est alors, un système qui offre une interface d'accès unifiée pour assurer l'interopérabilité entre différentes sources de données, en transformant par réécriture, les requêtes d'un utilisateur en sous-requêtes renvoyées aux sources de données les plus appropriées [Parent, 96]. Selon les travaux de [Wiederhold, 92] et [Inmon, 93], l'architecture d'un SID est basée sur l'utilisation d'un schéma global, qui fournit une vue réconciliée ou unifiée des sources locales. Ce schéma global consiste à consolider tous les schémas locaux des sources en un seul schéma global, qui sera utilisé pour supporter les requêtes. La figure suivante illustre l'architecture générale d'un SID.

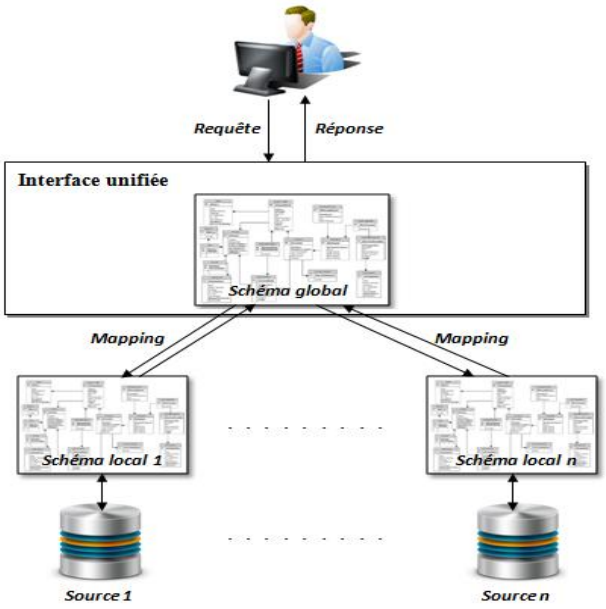


Figure 2.1. Architecture générale d'un système d'intégration de données [Wiederhold, 92]

À travers le système d'intégration de données, l'interrogation de sources hétérogènes est faite par les requêtes qui sont exprimées en termes de vocabulaire du schéma global. Il est alors nécessaire, d'identifier les correspondances (ou mapping) entre le schéma global et les schémas locaux.

Historiquement, Witold Litwin [Litwin, 85] a proposé un premier type des systèmes d'intégration de données, appelés Systèmes multibases. Ces systèmes permettent de gérer et stocker les multi bases de données relationnelles (*MRDSM* : Multi Relational Database Store and Management). Ils sont capables d'inter opérer les différentes sources de données hétérogènes sans une vue commune. L'accès à ces sources est assuré grâce à un langage commun tel que SQL.

Entre 1986 et 1989, Litwin et al [Litwin, 89] ont proposé à l'INRIA, un langage multibase de SQL appelé MSQL (Multi database Structured Query Langage), dont sa requête peut être formulée en calcul de prédicat d'ordre supérieur à 1, de faire l'optimisation en algèbre multi-relationnelle et elle peut produire une multi-table relationnelle. Trois nouvelles expressions ont été ajoutées à la syntaxe d'une requête SQL : CREATE MULTIDATABASE, ALTER MULTIDATABASE et DROP MULTIDATABASE.

À partir de MSQL, l'utilisateur peut exprimer ses besoins, envoyer autant de requêtes aux diverses sources et doit relier les différentes réponses aux requêtes formulées. Le Sybase est le premier SGBD commercial qui utilise les requêtes multibases [Litwin, 89].

Les systèmes multibases quoiqu'assurent un accès unifié via le langage multibase MSQL, ils ne permettent pas d'unifier la sémantique des données de différentes sources. À l'inverse des systèmes multibases, les systèmes fédérés où les bases de données sont appelées *Bases de données fédérées* (Federated database), assurent un accès unifié via une vue commune [Heimbigner, 85]. Cette vue est représentée par le *schéma fédéré* qui permet d'unifier les schémas des sources et de traiter leur hétérogénéité.

L'intégration de sources de données hétérogènes via un système fédéré consiste d'abord à traiter l'hétérogénéité sémantique au niveau de données et l'hétérogénéité syntaxique par la traduction des schémas, puis l'intégration des schémas pour créer un schéma global [Sheth, 90].

Néanmoins, plusieurs problèmes restent à résoudre à savoir, la conception d'une représentation commune des données partagées et l'adaptation aux différents systèmes distribués [Parent, 96]. En général, d'autres approches ont mis en amont pour l'intégration de données et qui devient nécessaire de les représenter en détail dans la section suivante.

3. Approches d'intégration de sources de données hétérogènes

Selon la localisation de données intégrées, deux principales approches d'intégration de sources de données hétérogènes ont été proposées [Parent, 96] : *approche virtuelle,* où les données restent dans leurs sources d'origine et *approche matérialisée,* où les données sont dupliquées dans un entrepôt de données. Les systèmes à base d'approche virtuelle sont

appelés systèmes de médiation et ceux qui sont basés sur l'approche matérialisée sont dits systèmes d'entrepôt de données [Calì, 13]. En effet, nous pouvons baptiser ces deux approches respectivement *approche de médiateur* et *approche d'entrepôt de données*.

3.1. Approche de médiateur

L'approche de médiateur est une approche virtuelle qui permet de donner l'impression à l'utilisateur d'interroger un système unique et homogène alors que les sources interrogées sont réparties, autonomes et hétérogènes [Sellami, 14]. Wiederhold [Wiederhold, 92] définit le concept de médiateur comme une couche logicielle qui exploite des connaissances afin que l'utilisateur peut accéder de manière transparente à différentes sources réparties et hétérogènes.

L'approche de médiateur consiste à offrir une interface unifiée d'interrogation via des requêtes plus signifiantes, exprimées en termes de vocabulaire du schéma global [Wiederhold, 92]. Ce dernier constitue une représentation virtuelle des données à travers l'exploitation des vues abstraites qui décrivent le contenu des sources.

Le traitement de requête initiale soumise par l'utilisateur est s'appuyé sur la réécriture de cette requête formulée en termes du schéma global en des plans de requêtes exprimées en termes de vue sur les sources de données [Vidal, 13]. La réécriture de requête joue un rôle primordial dans le processus d'intégration de données à base de médiateur, elle est fondée sur la manière d'identification des correspondances entre les éléments du schéma global et ceux de schémas locaux, en prenant en compte l'hétérogénéité structurelle et sémantique entre les schémas [Bouchou, 14].

Selon les travaux de [Vidal, 13] et [Calì, 13], pour exécuter les requêtes réécrites, il est nécessaire de les adapter selon le langage des sources à travers des adaptateurs (ou Wrappers) propres à chaque schéma de sources. Ces adaptateurs traduisent les plans de requêtes du langage de médiateur en un ensemble de requêtes compatibles au langage des sources puis ils les transmissent aux différentes sources pour réaliser leur exécution. De plus, les adaptateurs transforment les réponses aux requêtes en des réponses conformes au schéma global du médiateur. Ces réponses sont par la suite combinées au niveau du médiateur en une seule réponse homogène et cohérente.

3.2. Approche d'entrepôt de données

À l'inverse de l'approche de médiateur qui consiste à intégrer les données stockées dans leurs sources d'origine, l'approche d'entrepôt de données conçu à créer une base de données unique appelée entrepôt de données (*data warehouse*) à partir des sources de données et de l'interroger par des requêtes exprimées en termes de vocabulaire du schéma global de l'entrepôt [Inmon, 93]. Ce dernier se définit comme une collection de données intégrées, orientées sujet, non volatiles, historisées, résumées et disponibles pour l'interrogation, l'analyse et la prise de décision rapide [Inmon, 93]. En effet, les données issues des différentes sources de données sont dupliquées après modification éventuelle au sein du système d'entrepôt de données [Inmon, 00]. La figure suivante représente l'architecture du système d'intégration à base d'entrepôt de données.

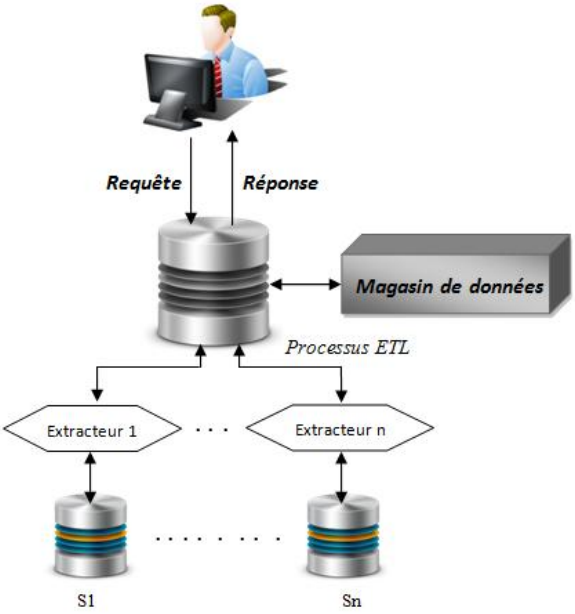


Figure 2.2. Architecture du système d'entrepôt de données [Inmon, 93]

L'intégration matérialisée de données est basée sur la construction d'entrepôt de données suivant le processus ETL (*Extract, Transform et Load*) qui permet d'*extraire* les données de sources via des extracteurs et d'appliquer diverses *transformations* aux données pour les nettoyer, les intégrer et les agréger; ensuite le *chargement* de données transformées dans l'entrepôt avec quelques changements aux données existantes [Vassiliadis, 03]. Cet entrepôt de données est alimenté depuis les données sources pour mettre à jour son contenu [Inmon, 93]. L'utilisateur peut accéder facilement aux données d'entrepôt comme une interrogation d'une seule source de données et leurs réponses sont retournées directement à l'utilisateur sans aucune transformation [Bellatreche, 03].

Ainsi, l'utilisateur peut interroger les magasins de données qui contiennent les vues matérialisées ou abstraites sur les données d'entrepôt afin d'effectuer des fonctions particulières telles que la fouille de données, des calculs prévisionnels et analyse des statistiques [Jarke, 13] [Khouri, 12]. La construction des entrepôts de données est un travail fastidieux ainsi que les données ne sont pas toujours fraîches et cela peut poser le problème d'incohérence entre les données stockées dans leurs sources d'origine et celles contenues dans l'entrepôt [Jarke, 13].

De nombreux projets de recherche sur l'intégration à base de l'approche entrepôt de données, les plus connus sont: **DWQ** (Foundations of Data Warehouse Quality) pour faciliter le choix des modèles et des structures de données selon des critères de qualité de service [Jarke, 97]. Les projets **WHIPS** (Warehouse Information Prototype at Stanford) [Labio, 97] et **SIRIUS** (Supporting the Incremental Refreshment of Information warehouses) [Gatziu, 98] ont pour but de gestion efficace de données de l'entrepôt.

3.3. Approche hybride

Une synthèse des travaux de [Abiteboul, 02] et [Boulçane, 08] nous a permis de présenter l'approche hybride d'intégration de données. L'approche hybride est une approche mixte qui combine à la fois l'approche de médiateur pour l'intégration des sources externes et l'approche d'entrepôt de données pour l'intégration de leurs données. Elle porte d'une part, sur la combinaison des approches de mappings (mises en correspondance) entre les schémas des sources et le schéma de médiateur, et d'autre part, sur la liaison entre le schéma de médiateur et le schéma d'entrepôt.

Dans un système d'intégration hybride, on peut y avoir plusieurs médiateurs pour intégrer les données d'entrepôt [Boulçane, 08]. Ces médiateurs peuvent être classifiés en deux familles : des médiateurs spécialisés offrent chacun une vue intégrée des sources de même modèle et un médiateur global pour intégrer les schémas partiels des médiateurs spécialisés. Ce médiateur global fournit un accès direct aux données d'entrepôt via une vue uniforme représentée par un schéma global de l'entrepôt. Un exemple de système hybride, le Xylème consiste à utiliser les documents XML issus du web et les intégrer à travers un mécanisme de vues entre les DTD concrètes des sources et des DTD abstraites montrées à l'utilisateur [Abiteboul, 02].

3.4. Synthèse

L'intégration de sources de données hétérogènes est un travail fastidieux et les approches proposées dans le cadre du traitement des problèmes d'hétérogénéité sémantique des sources de données sont classifiées en deux catégories : approches de médiateur et approche d'entrepôts de données.

L'approche de médiateur est basée sur la définition d'un schéma global décrivant de façon homogène et uniforme les schémas hétérogènes des sources de données. Les données restent dans leurs sources d'origine et une migration de requêtes de médiateur vers les sources de données est effectuée grâce aux adaptateurs. Cette migration consiste d'abord à réécrire ces requêtes en un ensemble de sous-requêtes équivalentes exprimées en termes de vues abstraites sur les sources. La migration de requêtes est fondée sur la définition d'une technique de mise en correspondance entre les éléments du schéma global et ceux des schémas locaux. Les adaptateurs effectuent l'interrogation effective des sources par la traduction de requêtes réécrites en termes de vues dans le langage propre à chaque source.

L'approche de médiateur permet une meilleure productivité au sein des entreprises qui exigent la manipulation directe de leurs bases de données autonomes, hétérogènes et distribuées. Elle présente l'intérêt de voir l'accès aux sources originales de données comme un accès à une seule source homogène. Néanmoins, quelques problèmes sont apparus à savoir, la réécriture des requêtes en termes de vues est très complexe, la conception des adaptateurs, la mise à jour du schéma global est très difficile dans le cas où le mapping entre schémas est faite d'une façon ascendante à partir des schémas locaux.

L'approche d'entrepôt de données où l'intégration est matérialisée par le fait que les données sont migrées vers un entrepôt qui joue le rôle d'une base de données unifiée et

homogène. La migration de données est effectuée après avoir accomplie les opérations d'ETL ; d'extraction, de transformation et de chargement de données dans l'entrepôt de données. Dans cette approche, le traitement de requêtes est très simple et faite directement sur l'entrepôt de données contrairement à l'approche de médiateur qui nécessite une traduction de requête au langage de sources. De plus, la création d'entrepôt de données permet d'assurer une meilleure intégration de données unifiées et non redondantes.

L'approche d'entrepôt de données simplifie l'optimisation et le traitement de requêtes en triant les données localement en fonction d'un seul schéma global. Par ailleurs, cette approche possède des problèmes tels que l'espace de stockage très grand à cause de duplication de données après modification éventuelle au niveau du système intégré, le processus d'ETL pour l'alimentation d'entrepôt est long et coûteux et exige d'espace mémoire pour effectuer les transformations, et la difficulté d'assurer la synchronisation entre les données d'entrepôt et les données originales stockées dans les sources hétérogènes.

Une étude comparative entre les deux approches d'intégration de données est illustrée dans le tableau suivant [Aggoune, 14b].

	Approche de médiateur (*Virtuelle*)	**Approche entrepôt de données *(Matérialisée)***
Principe	Interface unifiée	Copie unifié des sources
Performance	Défis principal	Bonne
Opérations sur les données	Néant	Les opérations d'ETL
Opérations sur les sources	Accès aux sources	Alimentation des entrepôts à partir des sources
Opérations sur les requêtes	Réécriture, Traduction	Néant
Opérations sur les résultats	Adaptation au schéma global	Néant
Evolution des sources	Souvent (Données fraîches)	Rarement (Données historisées et non volatiles)
Type de requêtes	Complexes et coûteuses	Simple et transactions longues
Complexité dans	Traitement de requêtes et conception des adaptateurs	Construction d'entrepôt de données, traitement de données
Traitement des problèmes d'hétérogénéité sémantique	Pendant le traitement de requêtes	Pendant la construction d'entrepôt

Table 2.1. Comparaison entre l'approche de médiateur et l'approche d'entrepôt de données

D'après la comparaison faite dans la table au-dessus, l'approche de médiateur est l'approche la plus adéquate pour l'intégration de données multimédias puisqu'elle ne nécessite pas des traitements complexes sur ce type de donnée contrairement à l'approche entrepôt de données où le processus ETL sur ces données est devenu plus complexe et lourd. De plus, les problèmes d'hétérogénéité sémantique sont étudiés pendant le traitement de requêtes qui est lui-même très coûteux, ce qui demande de trouver une solution pour remédier à ces problèmes et d'optimiser l'exploration de données avec un temps d'intégration court et un espace mémoire réduit.

Dans l'approche hybride, l'intégration de données est faite par la combinaison des deux approches précédentes ; une approche de médiateur pour intégrer les sources de données

externes puis la matérialisation de l'entrepôt conformément au schéma global du médiateur. On peut y avoir dans un système hybride plusieurs approches de mapping entre schémas. En effet, cette approche permet d'exploiter les avantages des approches de médiateur et d'entrepôt de données et essayer d'alléger leurs inconvénients.

Actuellement, le principal challenge des systèmes d'intégration est d'offrir une interface permettant de traiter les problèmes d'hétérogénéité sémantique pour l'exploration des sources de données multimédias et hétérogènes. Dans le cadre de notre travail, nous nous focalisons sur l'intégration de données via l'approche virtuelle du système de médiation en exploitant ses avantages par rapport les autres approches et nous apportons des solutions pour ses problèmes dans le cas de données multimédias. Dans cette optique, nous allons présenter dans ce qui suit l'architecture en couches d'un système de médiation.

4. Architecture d'un système de médiation

L'approche virtuelle par le système de médiation (ou le médiateur) propose une architecture très intéressante apportant une solution partielle aux problèmes d'hétérogénéité sémantique. Nous nous basons sur le travail de [Wiederhold, 92] pour présenter l'architecture d'un système de médiation. Cette architecture est constituée de structure en trois couches fondamentales: couche médiateur, couche adaptateurs et couche sources de données.

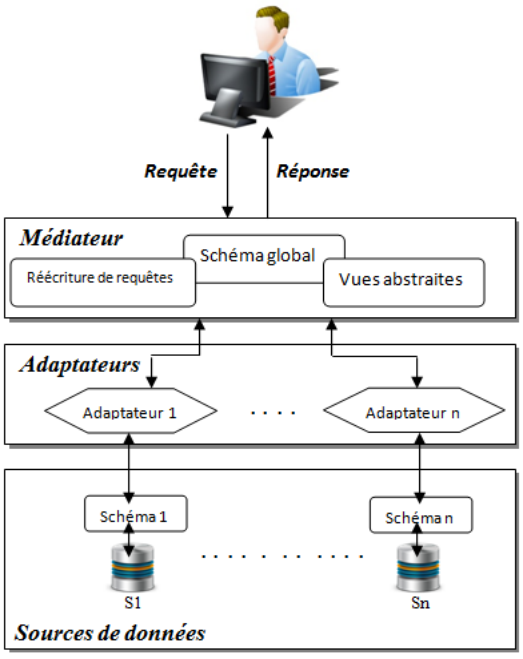


Figure 2.3. Architecture en couches d'un système de médiation [Wiederhold, 92]

Le système de médiation offre à l'utilisateur une interface d'accès unifiée aux sources de données hétérogènes. L'utilisateur peut donc poser ses requêtes exprimées en langage du médiateur et recevoir leurs réponses. Ces requêtes sont réécrites en termes des vues abstraites et par la suite elles sont transmises à la couche adaptateurs pour traduire à l'aide des adaptateurs les requêtes réécrites en un ensemble de sous-requêtes exprimées en langage de sources et envoyer cet ensemble à la couche sources pour l'exécuter. Les adaptateurs

reçoivent les réponses issues de la couche sources puis ils les combinent pour les envoyer au médiateur. Ce dernier reformate les réponses selon le vocabulaire du schéma global et il retourne finalement ces réponses à l'utilisateur. Nous détaillons dans le reste de cette section le fonctionnement de différentes couches du système de médiation.

4.1. Couche médiateur

La couche médiateur représente le point d'accès au système de médiation. Elle comporte un schéma global, un ensemble de vues abstraites et un module de réécriture de requêtes [Wiederhold, 92]. Le schéma global fournit un vocabulaire unique pour l'expression des requêtes des utilisateurs [Lenzerini, 02]. Il peut modéliser par un modèle relationnel, modèle orienté-objet, modèle logique, modèle sémantique, etc. Le choix d'un tel modèle dépend les besoins applicatifs et le type de problème à résoudre. Dans tous les cas, le modèle sémantique à base d'ontologie est le modèle le plus utilisé pour standardiser l'accès aux données et de traiter les différents problèmes d'hétérogénéité [Bouchou, 14]. Plusieurs approches ont été proposées dans la littérature pour la conception du schéma global, les plus connues sont : LAV (Local As View) et GAV (Global As View) (plus de détail en section 5.).

Le schéma global permet de décrire le contenu des sources de données par un ensemble de leurs vues abstraites [Sellami, 14]. Ces dernières sont alors exprimées par le même langage de description du schéma global. Elles permettent d'assurer l'intégration de sources par la réécriture de requêtes écrites en termes du schéma global en des plans de requêtes écrites en termes de ces vues.

La réécriture de requêtes est donc consiste à reformuler les requêtes posées par l'utilisateur selon le schéma global en des plans de requêtes exprimées en termes de vues sur les sources de données [Vidal, 13]. La réécriture de requêtes est une tâche plus complexe car elle doit prendre en compte la décomposition de requête et les problèmes d'hétérogénéité structurelle et sémantique entre les schémas. L'un des algorithmes le plus populaire pour la réécriture de requêtes est l'algorithme MiniCon qui permet de trouver à l'aide des comparaisons arithmétiques, les éléments similaires dans la requête initiale qui doit être conjonctive à celles de vues [Pottinger, 00]. Le résultat de réécriture de requêtes est un ensemble de plans de requêtes qui est par la suite soumis à la couche adaptateurs.

Généralement la couche médiateur repose sur :

- la définition d'un modèle ou schéma standard et uniforme;
- la description des sources de données via les vues abstraites;
- le mapping entre le schéma global et les schémas locaux;
- la réécriture de requêtes d'utilisateur en des plans de requêtes écrites en termes des vues abstraites des sources;
- la combinaison des réponses issues de la couche adaptateurs et effectuer éventuellement quelques opérations supplémentaires;
- et l'actualisation des sources soit par l'ajout ou la mise à jour d'une source.

4.2. Couche adaptateurs

La couche adaptateurs c'est la couche qui permet d'accéder au contenu des sources de données. Elle prend en entrée les plans de requêtes exprimées par le vocabulaire des vues abstraites suivant le schéma global et donne en sortie l'ensemble de requêtes traduites par les adaptateurs (*wrappers*) [Vidal, 13]. Ces derniers sont conçus à chaque source de données. Ils visent à traduire les entrées en requêtes exprimées en termes de langage de sources puis ils les envoient vers la couche sources de données pour faire leur exécution. La conception des adaptateurs est une tâche assez compliquée et elle est souvent basée sur les méthodes d'intelligence artificielle comme l'apprentissage par ordinateur [Vidal, 13].

Après avoir exécuté les requêtes, leurs réponses ont été transformées par les adaptateurs en des réponses adéquates au schéma global du médiateur. Les réponses sont ensuite renvoyées au médiateur qui se charge de les intégrer avant de les transmettre à l'utilisateur.

4.3. Couche sources de données

Cette dernière couche permet de représenter la grande quantité de données qui sont souvent stockées de façon distribuée dans des sources indépendantes [Vidal, 13]. Ces sources sont généralement conçues par différents concepteurs avec diverses méthodes et des besoins applicatifs différents. De plus, ces sources sont réparties sur plusieurs localisations : des ordinateurs différents, des partitions différentes, des réseaux sociaux, du fil RSS, des sites professionnels, des encyclopédies, etc.

Avec l'évolution des sources de données, celles-ci étant devenues hétérogènes selon différentes sortes [Beneventano, 13]: la langue utilisée (le multilingue), le média choisi (le multimédia), la structure et les problèmes d'hétérogénéité sémantique. En effet, le traitement des problèmes d'hétérogénéité sémantique est une des questions fondamentales à résoudre quand on veut explorer ces sources de données.

L'avantage de l'architecture du système de médiation est le pouvoir d'ajouter, de supprimer et de mettre à jour les sources de données, celles-ci étant souvent contenues des données multimédias. De ce fait, connaître les approches de mapping entre le schéma du médiateur (schéma global) et les schémas des sources, est indispensable pour une telle architecture.

5. Approches de mapping schéma global-schémas locaux

Dans un système d'intégration, des mises en correspondance ou mappings doivent être définies entre les relations du schéma global et celles des schémas locaux de sources de données à intégrer. Deux principales approches de mapping ont été proposées dans la littérature [Ullman, 97] [Halevy, 01]: l'approche GAV (Global As View) où les relations du schéma global sont exprimées en fonction des relations du schéma local et l'approche LAV (Local As View) où les relations des schémas locaux sont exprimées en fonction des relations du schéma global. D'autres approches hybrides sont basées sur la combinaison de ces deux approches à savoir, GLAV, BGLaV, BaV et HAV.

5.1. Approche GAV

L'approche GAV (Global As View) [Chawathe, 94] est la première approche utilisée pour assurer le mapping entre les éléments du schéma global et ceux des schémas locaux des sources de données à intégrer. Elle provient des systèmes fédérés où le schéma global est appelé schéma fédéré. Cette approche appelée aussi approche ascendante car les éléments du schéma médiateur (schéma global) sont exprimées en fonction des éléments des schémas de sources [Chawathe, 94]. En d'autre terme, les éléments du schéma global c'est-à-dire les relations et les attributs dans le cas du modèle relationnel, sont définis comme des vues (via l'expression CREATE VIEW) sur les éléments des schémas des sources à intégrer. En effet, la modification des sources (ajout, suppression ou mise à jour) implique également la modification des adaptateurs et du schéma de médiateur [Ullman, 97]. Cette tâche est généralement difficile à réaliser et elle représente un problème majeur dans la définition de mapping à base de GAV. En revanche, on trouve la même difficulté dans le cas inverse c'est-à-dire lors de modification du schéma global, les schémas locaux sont eux-mêmes doivent être modifiés en raison de dépendances entre ces schémas [Ullman, 97].

Pour illustrer le principe de cette approche, nous considérons un exemple d'intégration de deux bases de données relationnelles R1 et R2 comme suit :

Les schémas locaux de ces bases sont respectivement :

Article (ISSN, Titre_article, année)
Livre (ISBN, Titre_livre, année_édition)

La création de la vue globale "Publication" est faite via l'expression create view suivante :

```
CREATE VIEW Publication (Matricule, Titre, Année) AS
SELECT *
FROM R1 Article
UNION
SELECT *
FROM R2 Livre;
```

Finalement, le schéma global est la relation "Publication" qui contient trois attributs Matricule, Titre et Année. Pour l'interrogation de données, nous obtenons facilement une requête en termes des schémas des sources de données intégrées, en remplaçant les éléments du schéma global par leur définition.

Par exemple : « chercher les documents scientifiques publiés en 2016 ». Cette requête est exprimée en termes du schéma global comme suit :

```
SELECT *
FROM Publication
WHERE Publication.Année=2016;
```

Le module de réécriture de requêtes au niveau de la couche Médiateur permet de transformer la requête du schéma global en des plans de requêtes écrites en termes des vues

abstraites. En d'autre terme, la requête initiale est réécrite par l'union de deux requêtes, la première est appliquée à la base de données R1 et la deuxième sur la R2.

```
SELECT *
FROM R1. Article
WHERE R1.Article.année=2016
UNION
SELECT *
FROM R2. Livre
WHERE R2.Livre. année_édition=2016;
```

Parmi les systèmes utilisant GAV, nous pouvons citer HERMES [Subrahmanian, 95], TSIMMIS [Chawathe, 94] et MOMIS [Beneventano, 01].

5.2. Approche LAV

À l'inverse de l'approche GAV, LAV (Local As View) est une approche descendante qui suppose que le schéma global est préexisté et à partir de ce schéma en définit les schémas locaux des sources de données à intégrer [Levy, 96]. Il n'existe donc pas de correspondances directes entre les éléments globaux et les éléments locaux ce qui rend la réécriture de requête une tâche très difficile à atteindre [Halevy, 01]. En revanche, l'ajout de nouvelles sources est effectué facilement, il suffit de définir des mappings entre le schéma de cette nouvelle source et le schéma global [Levy, 96]. Prenons le même exemple décrit dans l'approche GAV, le schéma global SG suivant l'approche LAV est la relation Publication (Matricule, Titre, Année). À partir de ce schéma, nous définissons deux vues abstraites sur les bases de données R1 et R2.

```
CREATE VIEW Article (ISSN, Titre_article, année) AS
   SELECT *
   FROM SG. Publication;

CREATE VIEW Livre (ISBN, Titre_livre, année_édition) AS
   SELECT *
   FROM SG. Publication;
```

Les principaux systèmes développés autour de cette approche sont: OBSERVER [Mena, 00], PICSEL [Rousset, 02] et Infomaster [Genesereth, 97].

5.3. Approches hybrides

Au-delà des approches GAV et LAV, une fusion entre les deux permet de découvrir des approches hybrides comme celles de GLAV, BAV, BGLAV et HAV. Nous présentons brièvement ces approches par ordre chronologique, de plus ancien au plus récent:

- L'approche GLAV (Global Local As View) [Friedman, 99] consiste à mettre en place des vues abstraites de schéma global et les schémas locaux tout à la fois. En effet, la direction de mapping entre schémas est bidirectionnelle, c'est-à-dire, l'application de

mapping du schéma global vers les schémas locaux issus de l'approche LAV et vise versa (selon l'approche GAV).

- L'approche BAV (Both As View) [Boyd, 02] est une approche mixte qui combine l'approche LAV et GAV, elle emploie des mappings basés sur des transformations réversibles (transformation pathways) de schémas. Elle a été introduite dans le cadre de plusieurs projets comme le projet d'intégration AutoMed.
- Dans l'approche BGLAV (Both Global Local As View) [Xu, 04], le mapping est fait entre un schéma source et un schéma cible prédéfini. Les schémas sources peuvent être un schéma global du médiateur ou des schémas locaux des sources de données. Les inconvénients des approches LAV et GAV sont donc réduits.
- L'HAV (Hybrid As View) [Boulçane, 08] est une approche qui vise à combiner le meilleur de GAV et de LAV, elle utilise des médiateurs spécialisés entre le schéma global du médiateur global et les schémas locaux des sources de données à intégrer. Ces médiateurs spécialisés sont fondés sur le mapping schéma spécialisé-schéma local en se basant sur LAV pour mettre à jour facilement les sources de données (avantage de LAV) et l'utilisation d'un médiateur global au niveau supérieur des médiateurs spécialisés afin d'assurer le mapping entre le schéma global et les schémas spécialisés. Ce mapping est basé sur l'approche GAV pour faciliter la réécriture de requêtes.

5.4. Synthèse

L'intégration de données via un médiateur nécessite la définition d'une stratégie de mapping entre le schéma global et les schémas locaux de chaque source de données. Le mapping à base de l'approche GAV est caractérisé par la simplicité du processus de réécriture de requête. Ce processus consiste à remplacer les éléments du schéma global de la requête par leur définition et à nettoyer la requête locale par l'élimination de toutes les redondances due au remplacement des éléments. Cependant, l'ajout d'une nouvelle source peut engendrer des mises à jour complexes du médiateur.

Contrairement à GAV, LAV facilite la modification des sources de données à cause de l'indépendance entre le schéma global et les schémas locaux, par contre, la modification de schéma global peut être difficile à étudier au niveau des vues définissant les sources. De plus, la réécriture de requêtes est reconnue comme étant difficile et la seule solution vise à chercher les requêtes locales équivalentes de la requête donnée [Halevy, 01].

Les avantages des approches GAV et LAV ont été combinés dans des approches hybrides. Parmi ces approches, on trouve GLAV dont les vues sont représentées au niveau du schéma global et les schémas locaux. Cette approche ayant des caractéristiques proches de LAV en termes de l'indépendance du schéma global, la simplicité d'ajout de nouvelles sources et la complexité de réécriture de requêtes. L'approche BAV est une approche similaire à GLAV, mais elle est très coûteuse pour le traitement de requêtes. L'approche BGLAV est basée sur le mapping (schéma source)-(schéma cible) et cela d'une façon indépendante pour chaque source. La complexité de réécriture de requêtes s'en trouve donc réduite. Dans l'approche HAV deux types de médiateurs ont été définis, des médiateurs spécialisés pour assurer le mapping entre leurs schémas et les schémas locaux et un médiateur

global qui joue le rôle d'un médiateur principal pour faire l'intégration de données. Cette approche peut être utilisée aussi bien dans un système de médiation que dans un entrepôt de données. Quoique cette approche soit très flexible pour la mise à jour des sources et le traitement de requêtes, son utilisabilité dans un projet de recherche reste une perspective à développer afin d'évaluer correctement la performance d'un système d'intégration à base de l'approche HAV.

D'une manière sommaire, l'approche LAV est plus importante pour assurer l'extensibilité de sources de données et donne un meilleur résultat dans le cas où le schéma global n'est pas fréquemment modifiable. La réécriture de requêtes devra être traitée notamment dans le cas de données multimédias. Dans le cadre de notre travail, nous nous sommes intéressés à l'approche LAV pour effectuer le mapping dans un système de médiation dédié au traitement du problème d'hétérogénéité sémantique des sources de données multimédias. En effet, la section suivante décrit les différentes stratégies d'intégration virtuelle de données.

6. Stratégies d'intégration virtuelle du système de médiation

Au début, l'intégration virtuelle de sources de données a été conçue manuellement par un expert, puis il devient semi-automatique grâce à l'association d'un programme et par la suite l'automatisation complète de processus d'intégration.

6.1. Intégration manuelle

L'intégration manuelle est la première stratégie d'intégration qui consiste à assister un expert humain dans le processus d'intégration de données. L'expert permet d'effectuer les mappings entre les schémas et de gérer les problèmes d'hétérogénéité syntaxique et sémantique [Chawathe, 94]. Plusieurs systèmes ont été développés selon cette stratégie comme les systèmes multibases de données, systèmes fédérés, le système TSIMMIS [Chawathe, 94]. Par ailleurs, cette stratégie devient impraticable lorsque le nombre de sources de données à intégrer est important et de nature multimédia, ou lorsque les sources évoluent fréquemment.

6.2. Intégration semi-automatique

Le processus d'intégration manuelle a été amélioré dans le cadre du traitement automatique de problèmes d'hétérogénéité sémantique en utilisant une ontologie linguistique Wordnet pour fournir la sémantique au schéma global [Visser, 99]. Le travail de l'expert humain est donc concentré à la mise en correspondance entre le schéma global et les schémas locaux [Visser, 99]. De plus, il doit vérifier le fonctionnement de Wordnet dans le système d'intégration. Apporter une certaine automatisation dans le processus d'intégration permet une réduction du temps de réponse et du coût. Le système KRAFT est l'un des systèmes de médiation qui consiste à associer une ontologie linguistique aux éléments du schéma global et cela d'une façon explicite [Gray, 97].

6.3. Intégration automatique

L'intégration automatique est l'intégration les plus utilisées qui vise à libérer complètement l'expert humain dans le processus d'intégration [Wache, 01]. En effet, un programme informatique permet d'effectuer toutes les tâches du médiateur, à savoir, le mapping, la réécriture de requêtes, l'adaptation de requêtes et bien sûr l'exécution de requêtes. Un traitement automatique des problèmes d'hétérogénéité sémantique consiste à utiliser une ontologie de domaine décrivant les concepts et les relations sémantiques entre eux [Sultan, 13].

L'emploie d'ontologie dans un système d'intégration permet d'éliminer automatiquement les conflits sémantiques entre les sources. Le projet BUSTER est fondé sur l'annotation sémantique des sources données [Visser, 04].

Dans le cadre du traitement de problèmes d'hétérogénéité sémantique dans un système de médiation, l'intégration automatique de données devra s'appuyer sur l'utilisation d'une ressource sémantique, les plus connues sont les ontologies.

En effet, nous présentons dans ce qui suit les différentes approches de médiation sémantique à base d'ontologie.

7. Médiation sémantique à base d'ontologie

Ces dernières années, les problèmes d'hétérogénéité sémantique ont favoris l'émergence des systèmes d'intégration sémantique permettant de traiter ces problèmes, à savoir un entrepôt sémantique et un médiateur sémantique [Bellatreche, 13] [Wache, 01].

Par ailleurs, l'appariation du Web sémantique suscite l'utilisation des ressources sémantiques telles que les ontologies. Ces dernières facilitent l'interopérabilité entre plusieurs sources de données hétérogènes et permettent la communication entre utilisateurs via leur vocabulaire partagé [Studer, 98]. Grâce aux ontologies, les systèmes devraient intégrer des données dans un contexte sémantiquement riche.

Les ontologies sont vues comme un moyen de standardisation d'accès aux données et un modèle unifié de données. Leur utilisation dans un système de médiation améliore notablement sa qualité avec des problèmes réduits d'hétérogénéité sémantique [Bellatreche, 13].

Dans cette optique, nous présentons les approches de médiation sémantique à base d'ontologie. Wache et al. [Wache, 01] identifient trois principales approches :

- médiation sémantique à base d'une seule ontologie;
- médiation sémantique à base de multiples ontologies;
- et médiation sémantique par hybridation.

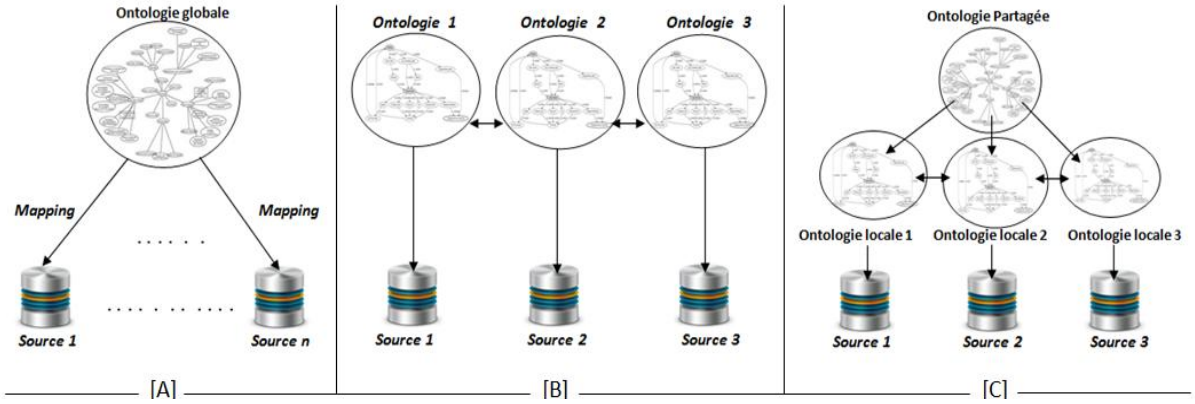


Figure 2.4. Les approches de médiation sémantique à base d'ontologie [Wache, 01]

7.1. Médiation sémantique à base d'une seule ontologie

La médiation sémantique à base d'une seule ontologie consiste à utiliser une ontologie de domaine pour la définition de schéma global du médiateur [Arens, 93]. Cette ontologie est appelée donc ontologie globale, qui a pour but d'assurer l'intégration de données et l'établissement des liens sémantiques entre les éléments de l'ontologie globale est ceux des schémas locaux [Wache, 01]. Elle fournit un vocabulaire commun et partagé pour l'expression des requêtes utilisateurs (ou requêtes de médiateur) et la description sémantique de contenus des sources. En effet, toutes les sources de données sont liées avec la même ontologie globale et leurs modifications impliquent pratiquement la modification de cette ontologie (cf. partie A, figure 2.4). Cependant, dans le cas de mapping via l'approche GAV, il devient difficile de mettre à jour l'ontologie globale ce qui n'est pas le cas dans LAV. Plusieurs systèmes sont basés sur cette approche à titre d'exemple SIMIS [Arens, 93] et PICSEL [Rousset, 02].

La médiation sémantique à base d'une seule ontologie est utile dans le cas où toutes les sources de données à intégrer possèdent presque les mêmes vues avec le même niveau de granularité [Ehrig, 04].

7.2. Médiation sémantique à base de multiples ontologies

La médiation sémantique à base de multiples ontologies représente une solution au problème de l'approche précédente où toutes les sources de données doivent partager la même ontologie globale [Mena, 00]. Or, ces sources sont souvent hétérogènes et ne possèdent pas les mêmes structures ni la même sémantique, avec des niveaux de granularité différents. L'approche à base de multiples ontologies consiste à attribuer à chaque source de données sa propre ontologie décrivant sa sémantique indépendamment des autres sources (cf. partie B, figure 2.4) [Sultan, 13]. De plus, ces ontologies peuvent être reliées entre elles par des liens sémantiques. Néanmoins, ces ontologies sont elles-mêmes hétérogènes en termes de concepts, de relations et de niveau de granularité ce qui rend les problèmes d'hétérogénéité sémantique non seulement au niveau des sources mais aussi au niveau de leurs ontologies. Par ailleurs, la mise à jour d'une source est très simple et elle implique la mise à jour de sa propre ontologie sans toucher les autres ontologies [Mena, 00]. Quelques exemples des systèmes à base de cette approche, OBSERVER [Mena, 00] et PIAZZA dans l'environnement P2P [Ives, 04].

7.3. Médiation sémantique par hybridation

Dans la médiation sémantique par hybridation, deux types d'ontologies ont été utilisés : une ontologie partagée (*Shared ontology*) qui est vue comme un schéma global du médiateur et un ensemble de ontologies locales (*local ontologies*) de la couche sources de données (cf. partie C, figure 2.4) [Gray, 97]. Chaque source de données à intégrer est associée à sa propre ontologie locale décrivant ses connaissances qui sont exprimées en termes du vocabulaire de cette ontologie [Visser, 04]. L'ontologie partagée représente le schéma global du médiateur, elle permet de définir un vocabulaire commun et partagé par les différentes ontologies locales afin d'éviter les conflits sémantiques entre les ontologies locales [Buccella, 05]. En revanche, toute modification ou ajout d'une source de données implique également la même opération sur l'ontologie locale associée à cette source et le mapping entre cette ontologie locale et l'ontologie partagée doit être amélioré [Buccella, 05]. Nous pouvons citer quelques systèmes d'intégration à base de cette approche : KRAFT [Gray, 97], COIN [Goh, 99] et BUSTER [Visser, 04].

7.4. Synthèse

Les ontologies jouent un rôle crucial pour le traitement des problèmes d'hétérogénéité sémantique dans un système de médiation intégrant des sources de données hétérogènes. Elles permettent d'expliciter la conceptualisation de connaissances qui sont généralement implicites et cachées [Albertoni, 11]. La médiation sémantique à base d'une seule ontologie donne un meilleur résultat dans des cas particuliers comme, l'utilisation des requêtes atomiques (simples), les sources à intégrer contiennent des documents n'auraient pas une structure précise comme les bases de données (données structurées) ou des sources de données structurées auraient une vue similaire du domaine avec le même niveau de granularité. Toutefois, cette approche dispose quelques inconvénients tels que le mauvais traitement de requêtes complexes et la difficulté d'effectuer l'intégration sémantique quand l'une des sources de données ayant des vues différentes et diverses niveaux de granularité. De plus, la modification dans la couche sources de données (ajout d'une source, suppression ou mise à jour) implique également une modification dans l'ontologie globale et dans le mapping avec les autres sources de données à intégrer.

La médiation sémantique à base de multiples ontologies est caractérisée par l'absence d'une ontologie commune entre les sources de données. En effet, chaque source de données est associée à sa propre ontologie ce qui rend les opérations de modifications des sources facilement manipulables d'une manière indépendante les unes des autres. Il suffit d'appliquer les mêmes opérations sur les ontologies, par exemple, l'ajout d'une source de données implique également l'ajout de son ontologie. Néanmoins, la liaison entre les ontologies est un travail fastidieux à cause de la subjectivité de concevoir une ontologie ; deux concepteurs différents ne produiront pas la même ontologie pour une même donnée. Ainsi, la difficulté de définir le mapping entre ontologies à cause de différentes agrégations et granularité dans les concepts d'une ontologie ce qui peuvent produire des problèmes d'hétérogénéité sémantique au niveau des ontologies.

Dans la médiation sémantique par hybridation, toutes les limites de l'approche à base d'une seule ontologie et celle à base de multiples ontologies ont été résolues. Ainsi, l'ajout d'une nouvelle source est effectué facilement sans aucune modification de mapping ou le vocabulaire de l'ontologie partagée. Le seul inconvénient de cette approche est la complexité de réutiliser une ontologie existante comme schéma global ; l'ontologie partagée devra être développée à partir de zéro pour définir un vocabulaire unifié et commun entre toutes les ontologies locales [Wache, 01]. De ce fait, il devra nécessaire de créer une ontologie standard pour supporter toutes les ontologies locales qui partagent le même vocabulaire de cette ontologie.

L'intégration sémantique de sources de données hétérogènes reste un défi pour le traitement des problèmes d'hétérogénéité sémantique qui se pose quelle que soit l'approche de médiation sémantique.

A partir de cette synthèse, notre travail se situe dans le cadre du traitement d'hétérogénéité sémantique en suivant l'approche hybride de médiation sémantique. D'autres problèmes auxquels nous avons dû faire face ont été de traiter d'une part, l'hétérogénéité sémantique dans les sources de données multimédias et d'autre part, l'hétérogénéité de modélisation de ces sources (le modèle relationnel et le modèle orienté-objet).

8. Intégration des sources de données multimédias

Les problèmes d'hétérogénéité sémantique sont non seulement dus à la diversité des sources de données mais également aux différents médias utilisés pour décrire la même donnée. Les données multimédias sont par nature hétérogènes ce qui engendrent beaucoup de problèmes (cf. chapitre 1, section 6.) qui rendent leur intégration un défi majeur dans le domaine d'intégration de données.

8.1. Problème d'intégration des sources de données multimédias

Très peu de travaux portent sur la problématique d'intégration de sources de données multimédias. Au début, l'intégration de ce type de données est effectuée grâce au système d'informations multimédias et hétérogènes qui consiste à définir un modèle d'accès unifié basé sur l'approche orienté-objet [Carey, 95].

D'autres travaux fondés sur l'exploitation des annotations et le contenu d'ontologie décrivant les données multimédias pour déterminer et intégrer les objets multimédias [Knoll, 98] [Petridis, 06]. Tous ces travaux quoiqu'assurent l'intégration de données multimédias, ils ne permettent pas de traiter le problème d'hétérogénéité sémantique de données multimédias.

L'intégration sémantique de données multimédias nécessite l'étude de nouvelle stratégie d'intégration en temps réel avec la prise en compte de la complexité croissante des données à intégrer. Nous nous focalisons seulement à la médiation sémantique à base d'ontologie pour l'intégration de données multimédias.

8.2. Médiation sémantique des sources de données multimédias

Déterminer et assurer l'intégration sémantique de sources de données multimédia reste un défi majeur. Récemment, une approche étendue a été proposée permettant l'intégration à la fois de sources de données traditionnelles et multimédias [Beneventano, 13]. La validation de cette approche est faite par le développement d'un système de médiation sémantique qui étend le système MOMIS qui a été proposé par la majorité de ces auteurs [Beneventano, 01]. Le système de médiation étendu est basé sur la fusion de deux systèmes existants : le système de médiation MOMIS et le système de gestion de contenu multimédia MILOS [Amato, 04]:

- Le MOMIS (Mediator envirOnment for Multiple Information Sources) est un système de médiation sémantique à base d'une seule ontologie, utilise la logique de description ODL-I3 comme langage commun pour décrire les schémas des sources de données structurées et semi-structurées [Beneventano, 01].
- Le MILOS est un système de gestion de contenu multimédia qui permet d'assurer le stockage et la recherche par contenu de tous types de documents multimédias [Amato, 04]. Il consiste à exécuter les requêtes sur ce type de documents qui sont généralement décrits par des métadonnées en langage XML.

Le système de médiation étendu possède les mêmes propriétés de MOMIS avec la capacité d'intégrer et d'interroger les données multimédias (documents et BDMM) via MILOS. Il utilise une ontologie globale 'SPDO' (Semantic Peer Data Ontology) comme schéma global pour décrire le contenu sémantique de données traditionnelles et multimédias dans un réseau de pairs [Beneventano, 13]. Le SQL-Like était le langage de manipulation de données multimédias et le langage ODL-I3 pour la gestion de requêtes globales de médiateur. De plus, le système adopte l'approche GAV pour le mapping entre SPDO et les sources de données à intégrer.

Le système de médiation sémantique présenté auparavant quoique capable d'intégrer les données multimédias, il ne garantit pas une meilleure performance à cause de manque d'expérimentations dans des scénarios réels. D'une façon générale, il n'existe pas des travaux relatifs dans ce contexte et l'intégration sémantique de données multimédias reste l'un des verrous scientifique à lever.

En conséquence, dans cette thèse, notre contribution est consacrée à proposer une nouvelle approche d'intégration sémantique des bases de données multimédias et hétérogènes.

9. Conclusion

Nous avons présenté dans ce chapitre les problèmes d'hétérogénéité sémantique des sources de données avec deux principales approches pour le traitement de ces problèmes : approche de médiateur où les données restent dans leurs sources d'origine et approche d'entrepôt de données, où les données sont dupliquées dans un entrepôt. Ces approches d'intégration des sources ont pour but d'assurer à l'utilisateur un accès unifié aux données sans besoin de connaître leurs sources d'origine. Plusieurs approches ont été présentées pour la mise en correspondance (mapping) entre le schéma global du système d'intégration et les

schémas locaux des sources de données à intégrer. Dans notre travail, nous nous sommes focalisés à l'intégration virtuelle de médiateur et l'approche LAV (Local As View) pour réaliser le mapping entre schémas et cela d'une façon automatique.

Par ailleurs, la médiation sémantique à base d'ontologie permet au mieux de traiter les problèmes d'hétérogénéité sémantique, en exploitant les liens sémantiques entre les éléments du schéma global qui est souvent une ontologie globale et les éléments des schémas locaux. Nous avons présenté également trois approches de médiation sémantique et nous avons mis l'accent sur la médiation sémantique par hybridation. Cette dernière utilise deux types d'ontologies ; une ontologie partagée au niveau du médiateur et des ontologies locales des sources de données à intégrer. La médiation sémantique de données multimédia reste un enjeu majeur dans les domaines d'intégration de données et de la recherche d'information sur le web. C'est dans ce contexte que se situe notre travail.

Les ontologies jouent un rôle crucial pour assurer une intégration sémantique de qualité. En outre, il est nécessaire de présenter dans le chapitre suivant, les concepts fondamentaux liés à l'ontologie ainsi que l'ontologie lexicale Wordnet que nous utiliserons dans notre travail.

CHAPITRE

03

ONTOLOGIES : UN ÉLÉMENT PRINCIPAL POUR LE TRAITEMENT D'HÉTÉROGÉNÉITÉ SÉMANTIQUE DE DONNÉES

Les ontologies jouent un rôle crucial pour traiter les problèmes liés à la sémantique. Il est donc nécessaire de présenter dans ce chapitre, les concepts de base des ontologies, leurs méthodologies de construction et leurs techniques d'alignement. Ce chapitre présente également le WordNet comme un moyen de désambiguïsation de sens des mots.

1. Introduction

Notre travail vise à proposer une approche qui permet de traiter le problème d'hétérogénéité sémantique pour l'exploration des sources de données multimédias. Nous voulons étudier ce problème d'hétérogénéité sémantique selon deux modes de représentation de données multimédias : les bases de données multimédias et les documents multimédias. Les ontologies sont souvent vues comme des ressources sémantiques les plus pertinentes pour définir du sens aux données. Elles jouent un rôle crucial pour désambiguïser le sens de mots contenant les données et les requêtes d'utilisateur. Il est donc nécessaire de prendre en compte de la sémantique de données et de requêtes via les ontologies pour l'exploration des sources de données multimédias, tout en traitant le problème d'hétérogénéité sémantique.

La médiation sémantique des bases de données multimédias, repose sur la construction puis l'utilisation des ontologies pour d'une part, définir un schéma unifié et partagé du médiateur et d'autre part, de décrire la sémantique de contenus des sources de données.

Dans le contexte de recherche d'information dans un corpus de documents multimédias, les ontologies sont utilisées pour assurer l'indexation sémantique de documents dont le but de définir un modèle de données d'un domaine particulier et des métadonnées pour décrire le contenu sémantique de documents multimédias.

Il est donc nécessaire de présenter dans ce chapitre les notions fondamentales liées aux ontologies avant la présentation de nos propositions dédiées au traitement d'hétérogénéité sémantique pour l'exploration des sources de données multimédias. Nous présentons également les approches et les méthodologies de construction d'ontologies avec les langages les plus utilisés pour la représentation et la manipulation des ontologies.

Avant de décrire les techniques d'alignement d'ontologie, il est nécessaire de présenter l'ontologie lexicale WordNet qui est utilisée pour trouver les relations lexicales entre termes telles que la synonymie et la polysémie. Elle permet de procéder des calculs de similarité sémantique entre termes, afin d'effectuer l'alignement entre deux ontologies ayant besoin de partager des connaissances entre elles.

Plusieurs techniques d'alignement ont été proposées et le choix d'une telle technique va permettre de résoudre les problèmes d'hétérogénéité sémantique des données. Nous terminerons ce chapitre par les outils les plus populaires pour, d'une part, la création et la gestion des ontologies et d'autre part, l'alignement des ontologies.

2. Définitions d'une ontologie

Le Web sémantique étend le web standard de W3C (World Wide Web Consortium) par l'ajout des métadonnées permettant de fournir une sémantique à l'information manipulée par l'utilisateur [Berners-Lee, 01]. Ces métadonnées ont été définies par des ressources sémantiques qui peuvent être des thésaurus, topic maps, ontologies, etc. Les ontologies représentent une technique primordiale du web sémantique, elles servent de vocabulaire standard pour le partage et la réutilisabilité des connaissances [Berners-Lee, 01].

L'origine du terme ontologie vient de la philosophie grecque (ontos=être et logos= études) plus précisément la métaphysique d'Aristote pour désigner l'étude de ce qui existe en général, c'est-à-dire tous objets reconnus comme existants [Welty, 01]. Par la suite, ce terme est utilisé dans différents domaines de l'informatique tels que, l'ingénierie des connaissances (IC), l'intelligence artificielle, la recherche d'information et l'intégration de données, dont le but d'exprimer et structurer les connaissances du domaine par un ensemble de concepts [Welty, 01]. En IC, Neches et al., [Neches, 91] ont donné une première définition de terme ontologie comme suit : « *Une ontologie définit les termes de base et relations comprenant le vocabulaire d'un sujet donné, ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire*». Par la suite, Thomas Gruber [Gruber, 93] a raffiné cette définition par la suivante : « *une ontologie est une spécification explicite d'une conceptualisation*» [Gruber, 93].

Autour de la définition de Gruber, plusieurs définitions ont été présentées dans la littérature, la plus élaborée est celle de Studer et al. [Studer, 98] qui combinent la définition de Gruber avec celle de Borst [Borst, 97]. Ce dernier, définit l'ontologie comme « *une spécification formelle d'une conceptualisation partagée* » [Borst, 97]. Studer et al ont défini alors, l'ontologie comme «*une spécification formelle et explicite d'une conceptualisation partagée*» [Studer, 98].

L'ontologie consiste à représenter les connaissances consensuelles d'une partie du monde réel par un modèle abstrait décrivant la hiérarchie des concepts et des relations sémantiques entre eux, c'est ce qu'on appelle « *la conceptualisation* » [Guarino, 09]. *La spécification formelle et explicite,* signifient que la conceptualisation est représentée par un langage interprétable par une machine et ses concepts et les contraintes liées à leur usage sont définis explicitement [Guarino, 09]. De plus, les connaissances consensuelles doivent être *partagées* par un groupe d'individus plutôt qu'elles soient utilisées par un seul individu [Guarino, 09].

3. Composants d'une ontologie

Une ontologie est donc une structuration des concepts sous forme hiérarchique. Pour présenter les différents composants d'ontologie, nous nous appuyons sur les travaux de [Guarino, 09] et [Teitsma, 14].

Les concepts sont un élément clé de l'ontologie en permettant la description explicite des connaissances. Ils sont également appelés classes, dont leurs noms doivent être non ambigus et non redondants. Si plusieurs termes désignent un même concept (ou classe), un seul d'entre eux sera choisi et identifié comme le terme préféré et les autres seront listés comme synonymes. Ainsi, on peut y avoir des sous-concepts qui héritent un super concept par le lien de subsomption is-a qui spécialise un super concept et généralise des sous-concepts, par exemple, le concept perishable-food hérite le concept food. De plus, les concepts peuvent être concrets (ex. food) ou abstraits (ex. flavor).

Chaque concept est défini par un ensemble de propriétés (ou slots) qui décrivent leurs caractéristiques intéressantes, par exemple, le concept microbe ayant comme propriétés :

family, binominal-name, discovered-by, date-of-discovery, length, et diameters. Ainsi, toutes les propriétés d'un concept doivent être dotées d'une valeur (ex. binominal-name: salmonella). De ce fait, des restrictions sur les valeurs des propriétés sont appelés facettes (facets), par exemple la propriété diameters est entre 0.7 et 1.5µm.

Par ailleurs, les concepts sont liés entre eux non seulement par le lien de généralisation/spécialisation *is-a* mais on peut y associer des relations prédéfinies et autres particulières à un domaine donné, par exemple: Part-of, Has-microbe, Has-Symptoms,…etc. Toutes ces relations induisent une organisation hiérarchique des concepts de l'ontologie. De plus, les propriétés peuvent aussi être organisées en hiérarchie de liens *is-a*. Un cas particulier des relations, appelées *fonction* dans laquelle un élément de la relation, le nième, est défini en fonction des n-1 éléments précédents.

D'un autre côté, un concept est décrit par un ensemble de individus (ou instances) véhiculent les connaissances, par exemple salmonella est un individu du concept Bacteria. D'autres composants complémentaires de la structure d'ontologie:

- *Restriction* : c'est une définition qui doit être vraie pour que ce qui est exprimé soit valable.
- *Règle*: c'est une affirmation logique décrivant des inférences possibles sous la forme antécédent>conséquent.
- *Axiome*: c'est une assertion toujours vraie, utilisée pour la définition des concepts et des relations.

En général, les règles et les axiomes sont utilisés pour vérifier la cohérence d'une ontologie, et aussi d'inférer de nouvelles connaissances. La figure suivante résume les différents composants d'une ontologie.

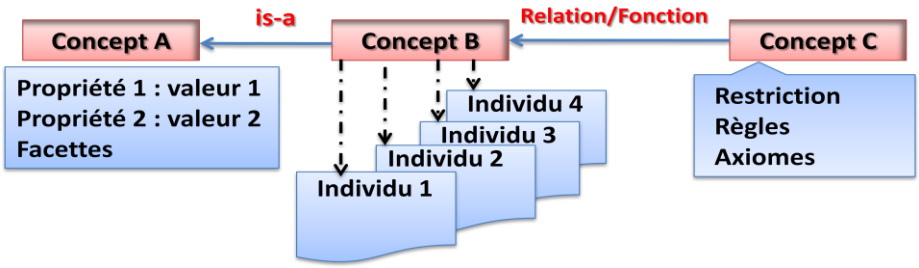


Figure 3.1. Les composants d'une ontologie

Les ontologies ont donc fourni un vocabulaire standard et partagé permettant d'exprimer la sémantique d'un phénomène ou d'une partie du monde réel. Elles permettent de lier les connaissances consensuelles du domaine à leur signification. De plus, les ontologies ont la capacité de partager et réutiliser des connaissances existantes.

4. Typologies des ontologies

Les ontologies sont vues comme une structuration hiérarchique des concepts d'un domaine. De ce fait, il existe plusieurs typologies des ontologies qui sont différenciées selon

des critères précis, les plus importants sont: l'objet de modélisation, le niveau de formalisation, le niveau de granularité et le type d'engagement.

4.1. Typologie selon l'objet de modélisation

Les ontologies peuvent être classifiées selon les objets que les modélisent. En effet, six classes ont été présentées qui sont:

- ***Les ontologies de représentation des connaissances :*** elles définissent un ensemble de primitives utilisées dans la représentation de langage des connaissances, telle que, l'ontologie *Frame-Ontology*[4] de langage de frame est représentée par classes, instances, facettes, propriétés, relations, restrictions, etc. [Van-Heijst, 97] [Gómez-Pérez, 99].
- ***Les ontologies de haut niveau :*** ou ontologies fondationnelles, elles définissent les connaissances de haut niveau et de sens commun via les concepts de niveau plus élevé comme les entités, les événements, les états, les processus, les actions, le temps, l'espace, etc. Elles sont fondées sur la théorie de l'identité et la théorie de la dépendance [Guarino, 97a] [Sowa, 95]. *OpenCyc*[5] est un exemple d'ontologie de haut niveau à vocation encyclopédique issue du projet Cyc.
- ***Les ontologies génériques ou méta-ontologies:*** elles sont proches des ontologies de haut niveau mais leurs concepts généraux sont moins abstraits et ils peuvent être réutilisés pour résoudre des problèmes génériques de différents domaines, par exemple l'ontologie *Mereology Ontology*[6] contenant des relations Associé-à [Van-Heijst, 97] [Gómez-Pérez, 99].
- ***Les ontologies de domaine:*** elles représentent les connaissances dédiées à un domaine particulier tels que: médecine, informatique, électronique, etc. [Mizoguchi, 00]. Plusieurs ontologies de domaines existent déjà, telle que l'ontologie *Geonames*[7] qui permet d'apporter de la sémantique aux données géographiques.
  - ***Les ontologies de tâches :*** elles définissent les connaissances portant sur des tâches et/ou des activités particulières dans les systèmes, telles que les tâches de conception et d'implémentation d'un logiciel [Mizoguchi, 00].
  - ***Les ontologies d'application :*** elles dépendent à la fois les ontologies de domaine et de tâches. Elles contiennent des connaissances de domaine pour une application particulière [Mizoguchi, 00].

La plupart des ontologies existantes sur le web sont des ontologies du domaine et c'est le type d'ontologie qui sera utilisé dans notre travail pour d'une part, traiter le problème d'hétérogénéité sémantique dans un médiateur d'intégration des bases de données multimédias et d'autre part, définir la sémantique de documents multimédias pour assurer l'indexation et la recherche de ce type de document tout en traitant le problème d'hétérogénéité sémantique.

[4] http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/frame-ontology/index.html
[5] http://sw.opencyc.org/
[6] http://ksi.cpsc.ucalgary.ca/KAW/KAW96/borst/kaw96doc.html
[7] http://www.geonames.org/ontology/

4.2. Typologie selon le niveau de formalisation

Selon le niveau de langage de formalisation des ontologies, Uschold et Gruninger ont distingué quatre types des ontologies [Uschold, 96]:

- ***Les ontologies informelles :*** elles sont écrites par un langage naturel non compréhensible par la machine.
- ***Les ontologies semi-informelles :*** elles sont écrites par un langage naturel structuré et limité.
- ***Les ontologies semi-formelles :*** elles sont écrites par un langage artificiel défini formellement.
- ***Les ontologies formelles :*** elles sont écrites par un langage artificiel contenant une sémantique formelle compréhensible par la machine.

4.3. Typologie selon le niveau de granularité

Les ontologies sont constituées des concepts qui peuvent être représentés en détail avec un niveau de granularité fine ou d'une façon généralisée avec un niveau de granularité large. En effet, les ontologies ont été classifiées en deux classes suivantes [Guarino, 97b]:

- ***Les ontologies de granularité fine :*** elles sont très détaillées, appelées aussi ontologies de grande taille. Elles possèdent un vocabulaire plus riche capable d'assurer une description détaillée des concepts pertinents d'un domaine ou d'une tâche.
- ***Les ontologies de granularité large :*** ou ontologies de petite taille, elles sont définies par un vocabulaire moins détaillé pour représenter des concepts de niveau précis.

4.4. Typologies selon le type d'engagement

Les concepts d'une ontologie disposent des caractéristiques permettant de les discriminer par rapport aux autres concepts. Ces caractéristiques nous ont permis ainsi de classifier les ontologies selon le type d'engagement : sémantique, linguistique, référentiel et opérationnel [Bachimont, 00].

- ***Les ontologies avec un engagement sémantique :*** elles définissent les concepts respectant les quatre principes de la structure hiérarchique qui assurent que chaque concept aura un sens (la sémantique). Ces principes sont: la communauté avec l'ancêtre, la différence par rapport à l'ancêtre, la communauté avec les concepts frères qui se situent au même niveau et la différence par rapport aux concepts frères. De plus, deux concepts sont sémantiquement identiques si et seulement si leurs interprétations à travers les quatre principes donnent un sens équivalent.
- ***Les ontologies avec un engagement linguistique :*** elles représentent un cas particulier des ontologies avec un engagement sémantique. Les ontologies linguistiques visent à définir le sens des mots et les relations lexicales entre ces mots. Elles sont basées sur une théorie linguistique du langage (les verbes, adjectifs, noms, adverbes, etc.). Le WordNet est un très bon exemple d'une ontologie linguistique qui représente les données lexicales de la langue anglaise sous forme d'un réseau sémantique. Il est utilisé pour la désambiguïsation du sens des mots.

- ***Les ontologies avec un engagement référentiel :*** elles se caractérisent par des concepts référentiels ou formels dont leur signification est définie par une extension d'objets. De ce fait, deux concepts référentiels seront identiques s'ils disposent la même extension.
- ***Les ontologies avec un engagement opérationnel :*** elles définissent par leurs concepts du niveau opérationnel ou computationnel. Ces concepts disposent des opérations permettant de générer des inférences. Ainsi, deux concepts opérationnels sont identiques s'ils possèdent le même potentiel d'inférence.

Les ontologies jouent un rôle crucial pour décrire correctement la signification des termes et améliorer l'exploitation des concepts par le partage et la réutilisabilité des ontologies existantes. Dans ce contexte, c'est une nécessité de présenter dans un premier temps les différentes approches de définition de la hiérarchie des concepts (ou classes) et dans un second temps les méthodologies de construction d'une ontologie.

5. Approches de définition de la hiérarchie des concepts

Le processus de construction d'une ontologie ou ingénierie ontologique consiste à collecter et représenter sous forme hiérarchique les concepts, les propriétés, les relations, etc. [Gruber, 93]. Cette structure doit offrir une signification explicite des concepts en fournissant des définitions objectives, cohérentes et extensibles sans que les nouveaux concepts influent sur les concepts existants [Gruber, 93].

De plus, l'ontologie devrait spécifier le moins possible la signification de ses concepts et son encodage afin d'assurer une bonne portabilité. Pour envisager les différentes méthodologies de construction d'ontologies, il devrait nécessaire de présenter les différentes approches de définition de la hiérarchie des concepts d'une ontologie. En effet, trois approches ont été proposées dans la littérature [Uschold, 96] [Noy, 01]: approche ascendante, approche descendante et approche hybride.

5.1. Approche ascendante

Dans l'approche ascendante, la création de la hiérarchie de concepts d'ontologie commence par la définition des concepts les plus spécifiques de la hiérarchie et se poursuit par la généralisation de ces concepts en concepts de plus haut niveau [Noy, 01]. Cette approche appelée aussi « de bas en haut » (*Bottom-up approach*) où son processus applique souvent une étude linguistique des structures de données existantes (documents, rapports, etc.) afin d'extraire des concepts pertinents du domaine et des relations entre eux [Uschold, 96]. L'approche ascendante se traduit par un niveau de détail très élevé qui rend difficile la localisation de points communs entre concepts voisins ce qui peut augmenter le risque d'incohérences.

5.2. Approche descendante

L'approche descendante vise à définir les concepts les plus généraux du domaine et s'approfondit dans la hiérarchie pour la définition des concepts plus spéciaux [Uschold, 96]. Cette approche appelée aussi « approche de haut en bas » (*top-down approach*). Le processus

de cette approche commence à analyser et à étudier les sources de données pertinentes dédiées à un domaine donné puis la modélisation des concepts de haut niveau qui seront raffinés dans des prochaines étapes. La définition de la hiérarchie des concepts par l'approche descendante est généralement réalisée manuellement par des experts de domaine et conduit à des ontologies de haut niveau, réutilisables et partageables.

5.3. Approche intermédiaire

L'approche intermédiaire (*Middle-out approach*) combine les deux approches précédentes (ascendante et descendante) dont le but de réduire leurs problèmes [Uschold, 96] [Noy, 01]. Les concepts de la hiérarchie se structurent autour de concepts saillants, ni trop généraux, ni trop spécifiques, puis ils sont généralisés par l'approche ascendante ou spécialisés par l'approche descendante [Uschold, 96].

Plusieurs recherches sont basées sur l'approche intermédiaire, la plus récente est celle de Ghosh et al. [Ghosh, 16]. Leur travail vise à construire une ontologie de référence au domaine de droit pénal et cela à travers la fusion de deux processus complémentaires : processus de modélisation conceptuelle basé sur l'approche descendante pour réutiliser les ontologies fondamentales et l'application de l'approche ascendante pour accomplir le processus d'apprentissage de l'ontologie à partir des ressources textuelles.

5.4. Synthèse

Trois principales approches existantes pour la définition de la hiérarchie des concepts d'ontologie. L'approche ascendante, appelée aussi bottom-up, consiste à définir les concepts de niveau plus détaillé et à généraliser vers des concepts plus généraux. Cette approche est utile pour la construction des ontologies d'application ou de domaines particuliers qui ne sont pas réutilisables. Ainsi, cette approche peut être utilisée comme support pour le raffinage et l'expansion des ontologies existantes en incorporant de nouvelles connaissances issues de textes.

L'approche descendante ou top-down définit avant tout, les concepts les plus généraux puis elle descend dans la structure hiérarchique vers les concepts les plus spécifiques. Cette approche assure un bon contrôle du niveau de détail. Toutefois, la définition des concepts généraux peut aboutir à choisir arbitrairement les concepts de haut niveau ce qui peuvent conduire à un risque de déséquilibre dans le modèle.

L'approche intermédiaire est une combinaison de ces deux approches. Elle est souvent, la plus facile à utiliser pour la plupart des constructeurs d'ontologies, étant donné que les concepts du niveau intermédiaire ont aptitude à être les concepts les plus descriptifs du domaine.

Aucune de ces trois approches de définition de la hiérarchie de concepts n'est meilleure que les autres, le choix d'une telle approche dépend fortement des objectifs de l'ontologie et du but à atteindre.

Dans cette thèse, nous appliquons l'approche top-down pour la définition de la hiérarchie des concepts d'ontologie du médiateur dédiée au domaine des risques alimentaires

et d'ontologie du modèle d'indexation personnalisée de documents multimédias dédiés au domaine de la recherche d'information.

6. Méthodologies de construction d'ontologies

Il existe dans la littérature plusieurs manières ou méthodologies de construction d'ontologies qui se diffèrent selon les phases de développement d'ontologie et les ressources utilisées [Drame, 14]. En effet, il n'existe pas une méthodologie communément admise, comme le cas dans le domaine des bases de données ; le choix d'une telle méthodologie dépend du phénomène du monde à concevoir et des besoins applicatifs [Drame, 14].

Dans ce cadre, il existe quatre principes fondamentaux de méthodologie de construction d'ontologie: la construction d'ontologies à partir de zéro, la construction d'ontologies à partir de texte, la construction d'ontologies par réutilisation d'ontologies existantes et la construction d'ontologies à base de crowdsourcing.

6.1. La construction d'ontologies à partir de zéro

Nous nous basons sur les travaux de [Lenat, 89] [Uschold, 95] pour définir la méthodologie de construction d'ontologies à partir de zéro (from scratch). Cette méthodologie représente le premier principe de l'ingénierie ontologique. Ce principe repose sur la collecte des connaissances nécessaires pour la construction d'ontologie à travers des réunions avec des experts, des interviews avec des spécialistes et les documentations existantes. De ce fait, la construction des ontologies est faite manuellement à l'aide des ressources humaines. Parmi les méthodologies qui suivent ce principe nous pouvons citer : CYC [Lenat, 89], Uschold et King [Uschold, 95], Gruninger et Fox [Gruninger, 95], METHONTOLOGY [Fernández-López, 97] et Noy et Mcguinness [Noy, 01].

Dans la méthodologie CYC [Lenat, 89], la construction d'ontologie est réalisée en deux étapes :

1) extraction manuelle de la connaissance implicite dans les différentes sources;
2) et utilisation des techniques de TALN et l'acquisition de connaissances pour générer de nouvelles connaissances à partir de celles acquises à l'étape précédente.

La méthodologie d'Uschold et King [Uschold, 95] a été proposée pour la construction de l'ontologie *Enterprise Ontology* décrivant les activités de l'entreprise. Elle fournit quatre étapes fondamentales :

1) identification de l'objectif et des utilisateurs prévus pour l'ontologie;
2) identification des entités nécessaire à la construction d'ontologie (concepts, relations, etc.);
3) évaluation et analyse de l'ontologie obtenue;
4) et la documentation facilitant la réutilisation et le partage de l'ontologie.

La méthodologie de Gruninger et Fox [Gruninger, 95] fait partie du projet TOVE (Toronto Virtual Enterprise). Elle est utilisée pour le développement des ontologies dans le domaine de l'entreprise. Cette méthodologie reste générique et aucune étape n'est pas décrite précisément. Elle consiste à exploiter les scénarios d'entreprises afin de les reformuler sous forme de questions auxquelles l'ontologie doit permettre de répondre. La spécification de la terminologie d'ontologie est effectuée par l'extraction des termes composant les questions.

METHONTOLOGY est une méthodologie qui a été proposé par l'équipe de Fernández-López du laboratoire de l'intelligence artificielle de l'Université polytechnique de Madrid [Fernández-López, 97]. C'est la méthodologie la plus largement utilisée, elle est inspirée de cycle de vie de développement de logiciel. La construction d'ontologie est faite grâce au processus en sept étapes suivantes: spécification des besoins, acquisition de connaissances, conceptualisation, intégration des ontologies existantes s'il est nécessaire, l'implémentation, évaluation et la documentation.

La méthodologie de Noy et Mcguinness [Noy, 01] est basée sur l'utilisation d'un processus itératif pour la conception d'ontologie tout au long de son cycle de vie. Ce processus comprend sept étapes suivantes :

1. *Spécification des besoins* : consiste à choisir le domaine qui va couvrir l'ontologie, ses objectifs et les utilisateurs qui vont utiliser et maintenir l'ontologie.
2. *Envisager une éventuelle réutilisation des ontologies existantes* : s'il possible de réutiliser des ressources existantes (ontologie, thésaurus, WordNet, etc.).
3. *Enumérer les termes importants dans l'ontologie* : construire une liste des termes importants pour concevoir l'ontologie.
4. *Définir les classes et la hiérarchie des classes :* la définition des classes et leurs relations afin de construire leur hiérarchie qui dépend le choix de l'approche de définition de la hiérarchie des concepts (ascendante, descendante ou intermédiaire).
5. *Définir les propriétés des classes :* pour chaque classe, dégager une liste de leurs attributs.
6. *Définir les facettes des attributs* : pour chaque attribut, définir les facettes décrivant la valeur du type, leurs cardinalités, leur type et les relations qu'elles entretiennent.
7. *Créer les instances* : consiste à associer à chaque classe un ensemble d'instances ou d'individus.

Bien que cette approche ne définisse pas l'encodage et l'évaluation d'ontologie, elle est très intéressante pour construire une ontologie en partant du zéro, car elle définit des règles très claires sur la définition des éléments d'ontologie (classes, propriétés, relations et instances) et le choix de sa structuration en hiérarchie.

Nous nous appuyons sur cette méthodologie pour concevoir nos ontologies de domaine avec un encodage et l'évaluation d'ontologie finale.

6.2. La construction d'ontologies à partir de texte

Le principe de construction d'ontologie à partir de texte a pour but d'automatiser certaines étapes de construction d'ontologies par l'utilisation des ressources textuelles et des outils du traitement automatique de la langue (TAL) [Drame, 14] [Rajpathak, 13]. Le principe de construction consiste à construire un corpus de documents textuels existants qui couvrent entièrement le domaine d'intérêt et l'application des outils linguistiques et statistiques [Rajpathak, 13]. Les outils linguistiques permettent d'analyser les relations lexicales entre termes, tandis que les outils statistiques sont fondés sur des calculs statistiques en utilisant des formules comme : TF (*term-frequency)* qui est le nombre d'occurrences d'un terme dans le document, et IDF *(Inverse of Document Frequency)* qui mesure l'importance d'un terme dans le corpus [Singh, 14]. Une phase de normalisation qui vise à construire la hiérarchie des concepts à partir des termes obtenus d'études linguistique et statistique. Par la suite, une formalisation du modèle conceptuel résultant de la phase précédente pour encoder par un langage formel [Rajpathak, 13].

TERMINAE[8] est une méthodologie et un outil de construction semi-automatique d'ontologies à partir de textes [Biebow, 99]. Elle est basée sur l'utilisation d'outils de TAL et repose sur un processus de six étapes : la constitution du corpus, l'analyse de ce corpus, l'étude terminologique, la normalisation, la formalisation et l'insertion dans l'ontologie.

L'outil TextToOnto[9] dispose des modules développés en Java permettant l'extraction des entités textuelles à partir de textes (termes, instances, relations, etc.) [Maedche, 00]. Il peut être intégré dans d'autres éditeurs d'ontologies comme dans le cas de l'éditeur KAON [Bozsak, 02].

Les méthodologies de construction d'ontologies à partir de textes se servent des outils TAL facilitant l'utilisation et l'exploration du corpus de documents. Néanmoins, ces méthodologies fonctionnent quand le corpus de documents est disponible.

6.3. La construction d'ontologies par réutilisation des ontologies existantes

La construction d'ontologie repose sur la réutilisation d'une autre ontologie préalablement construite par l'un des quatre principes existants [Maedche, 03]. Ce principe est connu sous le nom de réingénierie d'ontologies [Simperl, 09]. Ces méthodologies sont souvent automatiques ce qui facilitent la construction d'ontologies et réduisent son coût en termes de ressources humaines (experts, spécialiste, développeurs, etc.) survenues et de temps de construction. Elles peuvent combiner par exemple, une ontologie de domaine avec WordNet pour rendre explicite les relations extraites des ontologies existantes [Maedche, 03]. La méthodologie de Hepp et Bruijn [Hepp, 07] est un exemple de ces méthodologies. Elle consiste à limiter le travail de développeur d'ontologie à la validation de l'ontologie construite et de transformer des schémas et des taxonomies des concepts informels des ontologies ou des thésaurus existantes en des ontologies formelles légères.

8 http://lipn.univ-paris13.fr/terminae/index.php/Main_Page
9 https://code.google.com/p/text2onto/

D'autres méthodologies s'intéressent à la construction coopérative d'ontologies partagées par des groupes d'utilisateurs distants. Elles visent à mettre en place des registres, catalogues ou des métas ontologies pour stocker et localiser les ontologies existantes, et des techniques pour supporter la réutilisation et l'évolution distribuées de ces ontologies [Kozaki, 07].

6.4. La construction d'ontologies à base de crowdsourcing

Le principe de construction d'ontologie par les techniques de *crowdsourcing* ou externalisation ouverte, vise à exploiter les savoir-faire et l'intelligence d'un grand nombre de personnes pour réaliser les tâches effectuées par le développeur d'ontologie [Getman, 14] [Mortensen, 13]. La construction d'ontologie par un certain nombre d'utilisateurs facilite l'accomplissement des étapes du processus de construction et permet d'alléger les difficultés rencontrées pendant ce processus avec un temps très court par rapport les méthodologies précédentes [Mortensen, 13]. De plus, utiliser plusieurs utilisateurs pour construire une ontologie rend cette méthodologie un travail collaboratif ou au contraire un travail purement parallèle.

Getman et Karasiuk [Getman, 14] proposent une approche à base de crowdsourcing pour la construction d'ontologie de domaine du droit. Cette approche est appliquée sur 20 utilisateurs (étudiants en droit) ayant chacun une tâche de collecter les concepts propres à un sous-domaine de droit. L'ontologie finale a été évaluée par des experts du domaine qui sont estimés la couverture des concepts à plus de 90%.

Par ailleurs, le principe de crowdsourcing reste un nouveau principe de construction d'ontologie et plusieurs tests doivent prendre en compte pour vérifier son applicabilité et son utilité dans le domaine de l'ingénierie ontologique. Des limites principales que l'on peut noter pour ce principe sont, l'existence des synonymes entre les concepts qui ont été créés par différents utilisateurs et la difficulté de lier ces différents concepts pour construire l'ontologie finale.

Quelle que soit la méthodologie utilisée pour la construction d'ontologies, une fois l'ontologie a été construite, elle devra être évaluée pour vérifier que cette ontologie soit valide, adéquate et correcte. Poveda-Villalón et al. [Poveda-Villalón, 12] définissent six critères d'évaluation d'ontologie :

1) la satisfaction humaine de la suffisance des connaissances représentées dans l'ontologie;
2) la consistance logique d'ontologie qui peut être détectée via des raisonneurs;
3) le bon choix du langage ontologique pour représenter les connaissances;
4) la bonne structuration des concepts dans la hiérarchie;
5) la précision de la modélisation conceptuelle du monde réel;
6) et l'adaptation d'ontologie dans les applications qui lui sont destinées.

En revanche, les ontologies peuvent subir des évolutions qui mènent à des modifications nécessaires et une propagation de ces changements dans les autres ontologies qui en relient [Maedche, 03] [Stojanovic, 04]. Plusieurs techniques d'évolution d'ontologie ont été

proposées dans la littérature à titre d'exemple, la gestion de versions soit par le suivi des traces de changement de contenu d'ontologie [Ognyanov, 02] ou par la comparaison des différentes versions d'une même ontologie [Noy, 04].

7. Représentation et manipulation des ontologies

L'opérationnalisation ou le codage des ontologies est une étape très importante dans le processus de construction d'ontologies. Elle consiste à représenter l'ontologie sous un formalisme précis via un langage formel compréhensible par la machine [Fernández-López, 97]. Nous présentons dans cette section, d'une part, les formalismes de description et d'inférence de connaissances et d'autre part, les langages de représentation et de manipulation d'ontologie.

7.1. Formalismes de représentation et de manipulation des ontologies

Les connaissances sont exprimées à l'aide d'un formalisme de représentation d'ontologie qui peut être effectué selon deux modes de description : les logiques de description et les graphes conceptuels.

7.1.1. Les logiques de description

Les logiques de description (LD) sont le type de langage formel retenu par le Web sémantique pour représenter les ontologies et faire des inférences [Brachman, 85]. Elles représentent une réponse à la problématique entre l'expression des connaissances dans des formules (*expressivité*) et la capacité de calculer si une formule est un théorème dans un temps fini (*décidabilité*) [Brachman, 85].

Les LD s'inscrivent entre la logique propositionnelle et la logique des prédicats en permettant de représenter les concepts, les rôles et les individus par deux boîtes logiques [Patel-Schneider, 91]: la boîte terminologique (*T-Box*) qui contient les concepts et les rôles, et la boîte assertionnelle (*A-Box*) qui contient les individus ainsi que les règles et les contraintes qui s'appliquent aux concepts. Une ontologie décrite en logique de description est un couple (S, P) tel que S est la signature d'ontologie qui contient l'ensemble de concepts, de rôles et d'individus. Le P contient les axiomes de *T-Box* et les assertions de *A-Box.*

Les LD permettent aussi de manipuler les connaissances à travers un mécanisme d'inférence qui est souvent basé sur la relation de subsomption [Patel-Schneider, 91].

7.1.2. Les graphes conceptuels

Les graphes conceptuels ont été introduits par John Sowa en 1984 pour pouvoir d'expression de la langue naturelle avec le formalisme de la logique en s'inspirant des graphes existentiels et des réseaux sémantiques [Sowa, 84]. Un graphe conceptuel permet de représenter les terminologies et les assertions par les nœuds et les arcs respectivement où les concepts et les individus sont schématisés par des rectangles (nœuds) et les relations conceptuelles sont représentées par des ellipses (arcs) [Sowa, 84]. Il est possible d'effectuer des manipulations par le raisonnement en utilisant des algorithmes de la théorie des graphes ou en les transformant en formules de logique du premier ordre.

La figure suivante présente un exemple de graphe conceptuel pour la conception de l'expression suivante « une compagnie emploie une personne nommée Wang».

Figure 3.2. Exemple d'un graphe conceptuel

Quel que soit le type de formalisme choisi, il devient nécessaire d'encoder les modélisations par un langage de représentation et de manipulation des ontologies implémentées au sein d'un système informatique.

7.2. Langages de représentation et de manipulation des ontologies

Plusieurs langages de représentation et de manipulation des ontologies ont été proposés afin d'en faire des ontologies formelles autrement dites ontologies computationnelles, facilement exploitables et partageables par un grand nombre d'utilisateurs. SHOE (*Simple HTML Ontology Extension*) est un langage basé sur le langage HTML avec des extensions pour représenter des connaissances, il a été créé au sein de l'université de Marylan puis il a été adapté à la syntaxe de XML [Heflin, 98].

Le langage OML (*Ontology Markup Language*) est basé sur SHOE et XML, il permet de faire des représentations des graphiques conceptuels [Gómez-Pérez, 02]. XOL (*Ontology Exchange Language*) est basé sur OML, conçu pour la définition des ontologies du domaine de la bioinformatique [Gómez-Pérez, 02]. Un autre langage basé sur OML est le CKML (Conceptual Knowledge Markup Language) qui s'étend l'OML par la possibilité de représenter des contextes, de séquents, etc. [Kent, 99].

RDF (*Resource Description Framework*) a été créé par le W3C pour décrire des ressources Web [Lassila, 98], RDFS (*RDF Schema*) est une extension de RDF [McBride, 04]. L'OIL (*Ontology Inference Layer*) exprime les connaissances à travers les frames et le raisonnement par la logique de description, il a été combiné avec le langage DAML (*DARPA Agent Markup Language*) qui utilise la syntaxe RDF avec les primitives de représentation à base de frames, pour former le langage DAML+OIL [Horrocks, 02].

Finalement, OWL (*Web Ontology Language*) dérivé des langages RDF et DAML+OIL, il a été utilisé comme un standard par le W3C (World Wide Web Consortium) [Antoniou, 04].

Tous ces langages ont pour but de standardiser la représentation des connaissances. Nous présentons dans cette section les langages les plus importants et les plus utilisés pour représenter la sémantique des ressources. Ces langages sont : RDF, RDFS, OWL et le langage de manipulation des ontologies SPARQL.

7.2.1. RDF & RDFS

Le RDF (*Resource Description Framework*) est une recommandation du W3C développé pour décrire les ressources du web (données et métadonnées) à l'aide d'une structure à base d'XML [Lassila, 98]. Nous basons sur le travail de [Lassila, 98] pour décrire

RDF et RDFS. Le RDF représente une ressource (document, site web, partie de page web, page web, etc.) par le triplet de la forme *sujet-prédicat-objet*, sachant que le sujet décrit la ressource identifiée par son URI, le prédicat est une propriété ou relation entre deux sujets et l'objet est une donnée (littéral) ou une autre ressource.

Un ensemble de triplets est représenté par un graphe RDF où les nœuds peuvent être des sujets ou des objets et les arcs sont des prédicats. Le contenu du document RDF est proche de celui de XML, délimité par les balises <rdf:RDF et </rdf:RDF> et il fait référence à l'espace de noms RDF par l'expression suivante: rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#". Il inclut aussi des balises propres à RDF telles que: rdf:value, rdf:Resource, rdf:sequence, etc.

Le RDFS (RDF Schéma) étend le RDF par la possibilité de définir des hiérarchies de classes en utilisant l'espace de noms rdfs=http://www.w3.org/2000/01/rdf-schema# avec des relations de subsomption entre classes (rdfs:subClassOf) et des hiérarchies de propriétés par des relations de subsomption entre propriétés (rdfs:subPropertyOf) [McBride, 04]. Le document RDFS s'écrit toujours par les triplets RDF, en ajoutant des nouveaux mots-clés comme: rdfs:Class, rdfs:Property, rdfs:Domain et rdfs:Range.

Un exemple de document RDF Schéma donnant les définitions des classes Microbe, Virus et Food, sachant que la classe Virus est une sous classe de Microbe et la classe Food est liée avec la classe Microbe par la relation sémantique Has-Microbe.

```
<rdf: RDF
    xmlns: rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns: rdfs= "http://www.w3.org/2000/01/rdf-schema#"
<rdfs:Class rdf:ID="Microbe">
</rdfs:Class>
<rdfs:Class rdf:ID="Virus">
    <rdfs:subClassOf rdf:resource= "#Microbe"/>
<rdfs:comment> The Virus class is a sub class of Microbe.
</rdfs:comment>
<rdfs:Class rdf:ID="Food">
</rdfs:Class>
<rdfs:Property rdf:About="#Has-Microbe">
<rdfs:Domain  rdf:resource="#Food">
<rdfs:Range rdf:resource="#Microbe">
</rdfs:Property >
</rdf :RDF>
```

Bien que le RDFS soit un langage de balisage permettant de représenter l'ontologie par une hiérarchie des classes, il porte des difficultés pour représenter les ontologies ayant de fortes contraintes telles que: l'expression de la combinaison booléenne de classes, la définition de disjonction de classes, etc.

7.2.2. OWL

Le langage OWL (Web Ontology Language) a été recommandé par le groupe WebOnt de W3C dont le but d'améliorer le RDFS en définissant un vocabulaire plus complet pour la

description d'ontologies complexes [Dean, 04]. Il dispose des constructeurs facilitant la représentation détaillé des caractéristiques des classes, propriétés et relations, à savoir, les cardinalités, l'équivalence, la symétrie, la transitivité, etc. [Dean, 04].

OWL demeure le langage le plus représentatif d'ontologies, il est capable de faire des inférences à partir de sa version OWL-DL. Ainsi, OWL dispose trois sous-langages [McGuinness, 04]:

- ***OWL-Lite***: c'est la version légère et moins complexe d'OWL qui permet de représenter l'ontologie par des primitives simples à savoir, cardinalité 0 ou 1, la transitivité, l'intersection, etc. OWL-Lite est toujours décidable et donc facile à implémenter par un moteur d'inférence.
- ***OWL-DL***: langage basé sur la logique de description. Il étend le langage OWL-Lite avec la capacité d'exprimer des opérations sur les concepts telles que, union, disjonction, énumération, et la négation. Il permet de faire des raisonnements à travers des algorithmes d'inférence.
- ***OWL-Full***: c'est la version la plus complète et la plus expressive. OWL-Full dispose toutes les primitives du langage OWL avec la possibilité d'assurer le recouvrement des types: un concept peut aussi être un individu ou une propriété et réciproquement. Son inconvénient majeur vient de sa difficulté de faire des inférences.

À partir de ces trois langages d'OWL, nous peuvent conclure que plus le langage est expressif, moins il est décidable. D'une manière générale, le contenu du document OWL est défini par les triplets RDF avec une description très précise de la sémantique en utilisant des mots-clés préfixés OWL à titre d'exemple : owl:SymmetricProperty pour la symétrie des propriétés, owl:disjointWith permet d'affirmer que deux classes n'ont aucune instance commune et owl:sameClassAs sert à affirmer que deux classes sont identiques.

OWL est devenu un langage de standardisation des ontologies dans la communauté internationale de recherche du domaine de web sémantique [Suwanmanee, 05]. C'est le langage que nous utiliserons pour la représentation de nos ontologies de domaine. Ce langage est associé au langage SPARQL pour assurer l'interrogation et l'exploration des connaissances.

7.2.3. SPARQL

Le langage SPARQL *(Sparql Protocol And RDF Query Language)* est une recommandation du W3C pour l'interrogation de données du web sémantique représentées au format RDF [Pérez, 06]. La syntaxe du langage de requête SPARQL est inspirée de SQL avec l'absence de la clause FROM. Le SPARQL vise à manipuler les graphes RDF par un ensemble de commandes permettant de : rechercher, ajouter, supprimer, modifier, questionner, construire, décrire, etc. Ainsi, pour optimiser le stockage de données RDF volumineuses et faciliter leur manipulation, le SPARQL utilise une base de stockage des triplets, appelée TripleStore [Pérez, 06].

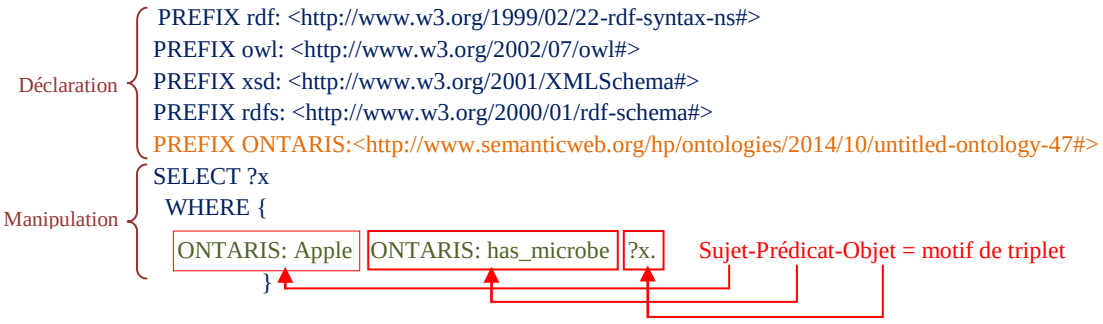
Une requête SPARQL est composée de deux principales parties [Pérez, 06]:

- la déclaration en format turtle: est une collection des espaces de noms (name space) ou encore des adresses IRI (Internationalized Resource Identifiers) qui généralisent et internationalisent les adresses URI. Chaque espace de nom est préfixé par le mot-clé PREFIX pour simplifier son utilisation dans la partie manipulation de graphe RDF.
- La manipulation de graphe RDF: cette partie comprend un ensemble de commandes à effectuer sur les motifs de triplet (triple pattern) qui sont comme les triplets RDF, à la différence majeure de possibilité d'exprimer les composants d'un triplet (sujet, prédicat et objet) par des variables.

La principale manipulation de données via des requêtes SPARQL est la recherche dans le graphe en utilisant deux clauses importantes [Quilitz, 08]:

- La clause **SELECT** : permet de présenter les variables que l'on désire voir figurer dans le résultat. Ces variables sont préfixées par le symbole '?' et séparées par des espaces. On peut également retourner toutes les variables du graphe par la commande select *;
- La clause **WHERE** : représente la requête proprement dite, délimitée par des accolades. Elle comporte un ensemble de motifs séparés par un point et un certain nombre des conditions telles que:
  - UNION entre deux conditions;
  - FILTER pour filtrer les réponses selon un critère donné (ex. FILTER ?age > 30);
  - ORDER BY pour trier selon une variable (ex. ORDER BY ?nom);
  - LIMIT pour limiter le nombre de réponses (ex. LIMIT 10),
  - OFFSET indique à partir de quelle position dans la séquence on démarre (ex. OFFSET 4);
  - Les fonctions d'agrégat: COUNT, SUM, MIN, MAX, AVG, etc.
  - La négation par le symbole '!', etc.

Exemple : chercher dans l'ontologie baptisée ONTARIS, les microbes qui peuvent coexister dans les pommes contaminées. Cette demande a été exprimée par la requête SPARQL suivante :



Dans cet exemple, la clause SELECT contient une seule variable x qui présente les instances de la classe liée avec l'instance Apple par la relation has_microbe. Ces instances sont de la classe Microbe. Le résultat d'une requête SPARQL est représenté sous une forme tabulaire qui peut contenir des données multimédias comme les images. La requête SPARQL peut utiliser dans la clause WHERE des conditions d'affichage optionnelle grâce au mot-clé OPTIONAL. Exemple : OPTIONAL {?ONTARIS:photomicrobe ?photomicrobe.} qui affiche tous les microbes, accompagnés de leurs photos, si la photo il y a.

Par ailleurs, pour éviter d'écrire un sujet d'un motif de triplet qui se répète dans plusieurs d'autres motifs d'une même requête, on utilise le caractère ';' entre ces motifs, par exemple, ONTARIS:Food ONTARIS:name "Egg" ; ONTARIS:type "Quail". La même chose dans le cas où on utilise plusieurs fois le même prédicat, on sépare les objets par le caractère ',', par exemple, ONTARIS:Bacteria ONTARIS:name "Escherichia-coli", "Pseudomonas".

Concernant les autres types de manipulation de données, le SPARQL permet d'effectuer l'ajout et la suppression de données par les commandes INSERT et DELETE respectivement. Pour la mise à jour d'une donnée est réalisée par une requête DELETE suivie d'une requête INSERT [Buil-Aranda, 13].

Par ailleurs, le SPARQL possède des opérations de gestion de graphes RDF comme : la création d'un graphe (CREATE GRAPH), le chargement (LOAD), l'effacement (CLEAR GRAPH), l'élimination (DROP GRAPH), le copier des graphes (COPY GRAPH… TO GRAPH), le déplacement d'un graphe dans un autre (MOVE GRAPH… TO GRAPH) et l'ajout des graphes (ADD GRAPH… TO GRAPH).

D'autre types de requêtes SPARQL permettent de gérer les graphes RDF qui sont [Quilitz, 08] [Buil-Aranda, 13]:

- **Requête ASK :** consiste à vérifier l'existence ou non d'une solution qui satisfait les conditions. Par exemple : ASK WHERE {ONTARIS:Food ONTARIS:name "Bread"}. Le résultat de cette requête est donc du type booléen (True ou False);
- **Requête CONSTRUCT:** vise à construire un graphe RDF solution à partir d'un graphe existant et selon un certain nombre de conditions. Cette requête est proche de création des vues (CREATE VIEW) à partir d'une requête SQL;
- **Requête DESCRIBE:** retourne un seul graphe RDF décrivant les ressources existantes;

Le langage SPARQL est souvent utilisé conjointement avec des langages de programmation comme Java pour manipuler des données RDF dans des applications. Il dispose de nouvelles fonctionnalités appropriées pour être utilisé comme un langage de requête commun à travers un moteur d'interrogation et d'inférence le plus connu est Jena[10]. Le SPARQL est moins expressif que SQL, en raison du manque de support de toute forme d'agrégation. Dans cette optique, nous nous sommes basés sur l'utilisation de langage SPARQL comme langage standard et unifié pour l'intégration sémantique des sources de données hétérogènes et multimédias.

[10] http ://jena.apache.org

Les ontologies représentent un moyen de définir les connaissances des sources de données et cela sous forme des métadonnées. Elles sont très utilisées pour décrire la sémantique de données notamment le type texte. De plus, pour représenter correctement la sémantique des données, il est nécessaire de vérifier leur aspect lexical. Dans ce contexte, plusieurs propositions majeures ont été apportées, la plus connue est l'ontologie lexicale WordNet que nous devrons la présenter dans la section suivante.

8. Le WordNet

Dans le cadre de représentation de connaissances, le domaine de l'IA (intelligence artificielle) a pour but d'identifier les relations conceptuelles reliant des concepts qui décrivent les connaissances [Welty, 01]. Ces concepts permettent de définir un ensemble de mots ou des termes relient entre eux par des relations lexicales [Miller, 90].

Définir les relations lexicales entre mots est l'un des objectifs du domaine linguistique. Les principales relations lexicales entre mots sont [Miller, 90] [Nefzi, 15]:

- *Polysémie :* la polysémie représente différents sens d'un même mot. Par exemple, le mot feuille peut indiquer les feuilles des arbres ou les feuilles d'un livre.
- *Synonymie :* différents mots ayant le même sens. Par exemple, les mots : content, joyeux, heureux permettent de désigner une personne qui éprouve du bonheur en raison de circonstances agréables et satisfaisantes.
- *Métonymie :* phénomène par lequel un concept est désigné par un terme désignant un autre concept qui lui est relié par une relation nécessaire. Par exemple, boire un verre signifie boire un verre d'eau et ce n'est pas le verre lui-même.
- *Hyponymie/Hyperonymie :* c'est la relation de généralisation/spécialisation is-a, sachant que l'hyponymie indique le rapport d'inclusion entre des unités lexicales les plus spécifiques aux plus générales. L'hyperonymie est l'inverse de l'hyponymie (relation entre une unité lexicale plus générale avec d'autres plus spécifique). Par exemple, le chien est dans un rapport d'hyponymie avec carnivore, animal, etc.
- *Méronymie/Holonomie :* c'est une relation partie-de, par exemple la poignée est une partie de la porte, alors la poignée est un méronyme de la porte. Les holonymes sont presque les inverses des méronymes.
- *Antonymie :* c'est la relation entre mots ayant un sens contraire à celui d'un autre, par exemple sympathique et hostile, laideur et beauté.
- *Toponymie :* c'est la relation d'hyponymie dans les verbes.

Ces différentes relations lexicales ont été utilisées dans une base de données lexicale de la langue anglaise appelée WordNet qui a été développée depuis 1985 par Georges Miller et ses collègues au laboratoire de sciences cognitives de l'université de Princeton [Miller, 90].

Le WordNet a été initialement conçu pour tester les déficits lexicaux dans des expériences de psychologie cognitive, par la suite, il a été utilisé dans le domaine de TALN et la recherche d'information sémantique afin de désambiguïser le sens des mots contenant un document ou une requête [Miller, 95].

Le WordNet représente les lexèmes par une structure hiérarchique porteuse de sens, en fait, il est considéré comme une ontologie lexicale où chaque lexème est décomposé en une collection de concepts lexicaux qui peuvent être des noms, des verbes, des adjectifs et des adverbes. Ces concepts sont représentés par un ensemble de synonymes regroupé en SynSets (Synonym Set) [Miller, 95].

La figure suivante présente un exemple du mot Thesis avec deux synset sous le Wordnet Search[11]- 3.1.

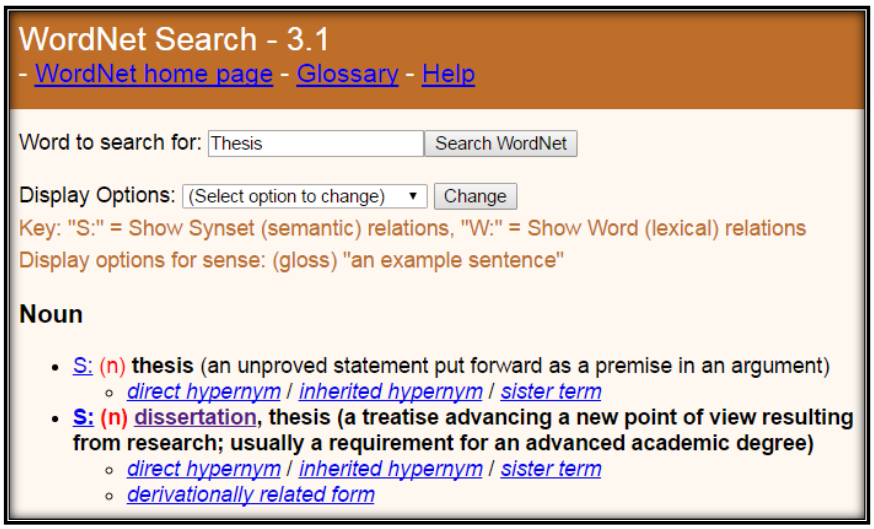


Figure 3.3. Recherche du mot Thesis via WordNet Search – 3.1 [2]

Un synset est représenté par une description du sens qu'il représente (gloss) avec des exemples illustratifs [Miller, 95]. Les Synsets sont reliés entre eux par des relations sémantiques et lexicales afin d'offrir une grande couverture lexicale. De ce fait, un mot ayant plusieurs sens doit appartenir à plusieurs synsets.

L'ontologie lexicale WordNet appelée aussi réseau lexical dans la mesure où les synsets représentent les nœuds et les relations entre synsets sont les arcs des nœuds.

Un exemple du nom Food qui dispose sous Wordnet 3.1, trois (03) sens différents avec leur description. Ces trois sens sont numérotés food#1, food#2 et food#3.

[11] http://wordnetweb.princeton.edu/perl/webwn

Le tableau suivant décrit quelques concepts liés au mot Food selon ses différents sens.

	Food		
	food#1 nutrient	***food#2 solid food***	***food#3 food for thought, intellectual nourishment***
Description	Any substance that can be metabolized by an animal to give energy and build tissue	Any solid substance (as opposed to liquid) that is used as a source of nourishment	Anything that provides mental stimulus for thinking
Hyponymes	▪ yolk#2, vitellus#1 ▪ comfort food#1 ▪ comestible#1, edible#1, victuals#3, ▪ feed#1, provender#1	▪ leftovers#1 ▪ fresh food#1 ▪ produce#1, green goods#1, garden truck#1 ▪ breakfast food#1	▪ pabulum#2
Méronymes	▪ allergen#1 ▪ pyrogen#2, pyrectic#1 ▪ adulterant#1, adulterator#1 ▪ digestive#1	None	None

Table 3.1. Extrait des synsets du mot Food via Wordnet Search-3.1 [2]

Le WordNet 2.1[12] répertorie plus de 115 000 synsets. Il est distribué sous une licence libre et bien documentée et il est gratuitement téléchargeable à l'URL suivante : https://wordnet.princeton.edu/wordnet/download/.

La dernière version de WordNet est 3.1 a été distribuée en avril 2013 [Finlayson, 14]. De plus, le projet EuroWordnet a permis la constitution des ressources Wordnet pour plusieurs langues: Anglais, Hollandais, Italien, Espagnol, Allemand, Français, Tchèque et l'estonien. Par exemple le WOLF (WOrdnet du Libre en Français) est une ressource linguistique généraliste libre pour le Français issue d'une traduction automatique de Wordnet [Sagot, 08]. Le BalkaNet pour les langues Roumain, Bulgare, Turque, Slovène, Grec, Serbe [Bond, 12]. Plusieurs autres ressources linguistiques ont été développées à partir de WordNet, à savoir, WordNet-Affect pour la représentation lexicale de connaissances sur les affects [Strapparava, 04], SentiWordNet pour le sondage d'opinion [Esuli, 07], WordNet Domains dédié aux domaines [Kolte, 08], etc.

Dans le cadre de notre travail, nous utilisons une ontologie de domaine pour représenter la sémantique de données hétérogènes et le WordNet pour la désambiguïsation de sens des mots qui offre une grande couverture lexicale. Le WordNet est utilisé ainsi pour détecter les relations lexicales lors de l'appariement entre les concepts de l'ontologie partagée du médiateur et les ontologies locales des sources de données.

En effet, la mise en correspondance ou l'appariement entre les concepts de deux ontologies est effectuée via une technique d'alignement. La présentation de ces techniques d'alignement d'ontologies sera l'objet de la section suivante.

[12] https://wordnet.princeton.edu/

9. Techniques d'alignement d'ontologies

Avant de présenter les techniques d'alignement d'ontologies, il est nécessaire de définir la notion d'alignement. Euzenat et Shvaiko ont donné la définition suivante: « *L'alignement d'ontologies est le processus de mise en correspondance ou appariement sémantique des entités qui les composent* » [Euzenat, 07]. Le processus d'alignement d'ontologies prend en entrée les ontologies à aligner et donne en sortie un ensemble de correspondances reliant les entités qui composent les ontologies alignées [Rahm, 11]. Ce processus est basé sur le calcul de similarité pour chaque couple d'entités afin de comparer tous les entités des ontologies alignées [Euzenat, 07]. En effet, selon le couple d'entités à comparer nous distinguons trois grandes catégories des techniques d'alignement : techniques terminologiques, techniques linguistiques et techniques structurelles.

9.1. Techniques terminologiques

Une synthèse des travaux présentés dans [Rahm, 11] [Shvaiko, 13] et [Cheatham, 13] permet de décrire les techniques terminologiques. Ces techniques sont basées principalement sur le calcul de similarité entre les chaînes de caractères composant les entités textuelles (noms, métadonnées sur les noms, étiquettes, commentaires, etc.).

L'alignement terminologique de deux ontologies revient à définir un seuil à une mesure de similarité entre les labels désignant deux concepts ou deux relations. Si la valeur de similarité est supérieure ou égale au seuil alors ces termes sont équivalents sinon ils sont distincts. La valeur de similarité doit être comprise entre 0 et 1. La valeur 0 indique qu'il n'y a pas une similarité entre concepts, tandis que la valeur 1 indique une similarité. Il existe plusieurs mesures calculant la valeur de similarité telles que [Stoilos, 05]: la similarité de Jaccard, la similarité de Jaro, la distance de Levenstein, etc.

Les techniques terminologiques sont efficaces dans le cas où les chaînes de caractères ne sont pas trop longues. Néanmoins, elles donnent un mauvais résultat quand les concepts à comparer sont sémantiquement proches et quand leurs noms sont différents (les synonymes).

9.2. Techniques linguistiques

Selon les travaux de [Safar, 09] et [Musen, 15], les techniques linguistiques permettent de remédier les problèmes des techniques terminologiques par l'association d'une ressource linguistique telle que dictionnaire, WordNet, afin d'indiquer les concepts sémantiquement similaires. La similarité entre concepts est faite en exploitant les relations lexicales existantes dans le WordNet par exemple. De plus, deux concepts sont similaires s'ils partagent la majorité de leurs synsets. Le WordNet est capable de calculer automatiquement la similarité sémantique entre concepts en utilisant des librairies et le package WS4J pour le langage de programmation Java.

9.3. Techniques structurelles

A partir les travaux de [Elbyed, 09], [Fellah, 08] et [Zghal, 07] nous présentons les techniques d'alignement structurelles. Les techniques d'alignement structurelles consistent à exploiter les informations relatives à la structure d'ontologies.

La similarité sémantique entre deux ontologies peut être basée sur la position des entités dans leurs hiérarchies ainsi si deux entités de deux ontologies sont semblables, leurs voisins le sont également d'une certaine façon. Cependant, cette mesure peut rencontrer quelques difficultés dans le cas où les ontologies à aligner ayant de niveau de granularité différent.

De plus, il est possible d'appliquer une mesure de similarité entre deux concepts par l'utilisation de leur structure interne qui décrite par l'ensemble d'informations incluent les restrictions et cardinalités sur les attributs, les instances, etc.

A partir des instances d'une classe on peut y déduire une nouvelle technique d'alignement appelée *Alignement extensionnelle* où chaque instance peut être représentée par un vecteur de noms et/ou de valeurs et des calculs de similarités entre vecteurs permettent de comparer les instances et donc les concepts des ontologies.

Des techniques sémantiques pour l'alignement d'ontologie s'intéressent à l'utilisation d'une troisième ontologie intermédiaire ou encore une ontologie commune telle que SUMO[13] (*the Suggested Upper Merged Ontology*) qui va permettre de découvrir les ambiguïtés de différentes significations possibles des termes de deux ontologies à aligner [Jean-Mary, 09]. De plus, les ontologies qui sont formalisées par les logiques de description (LD) peuvent être alignées en exploitant l'une des caractéristiques de LD telle que la déduction pour déduire la similarité entre deux entités [Baader, 99].

L'alignement d'ontologie s'impose comme une solution, pour permettre l'interopérabilité et le partage des sources de données hétérogènes. Il représente également un processus fondamental dans notre travail afin d'assurer l'intégration de données. La section suivante dresse quelques outils de manipulation et d'alignement d'ontologies.

10. Outils de manipulation d'ontologies

Cette section est consacrée à la représentation des outils les plus utilisés pour, d'une part, la création et la gestion des ontologies et d'autre part, l'alignement d'ontologie.

10.1. Outils d'édition d'ontologies

Quel que soit le domaine d'application des ontologies et le langage de programmation utilisé, des outils de développement d'ontologies ont été proposés pour aider l'utilisateur à créer et gérer facilement les ontologies. Les outils les plus utilisés sont : Protégé 2000, SWOOP et OntoEdit.

- ***Protégé 2000*** : l'éditeur Protégé 2000[14] est un logiciel open source pour la création et la gestion des ontologies, développé à l'université de Standford [Noy, 00a]. Il est basé sur le modèle des frames pour la construction des ontologies qui sont composées d'une hiérarchie des classes relient entre elles par des relations de subsumption *is-a* et des relations définies par le développeur. Une classe est décrite par son nom, ses propriétés (slots), ses facettes, ses fonctions et un ensemble de ses instances. Le Protégé 2000 ou encore Protégé-OWL, supporte plusieurs langages de représentation d'ontologie comme

[13] http://www.ontologyportal.com
[14] http://protege.stanford.edu/

RDF, RDFS, OWL, etc. [Noy, 00a]. Il est aussi un Framework extensible grâce à la disponibilité des plugins notamment pour la visualisation en mode graphique les ontologies, Jambalaya, TGViz, OntoViz, Graphviz, etc. [Knublauch, 04]. Grâce aux plugins de Protégé, il est possible d'importer des ontologies écrites dans différents langages d'ontologies (RDFS, OWL, DAML, OIL). Il dispose de raisonneurs (moteurs d'inférence) comme Racer, Fact++, Hermitt, Pellet [Noy, 00a] [Knublauch, 04].

- **SWOOP**: est un éditeur d'ontologies open source, produit par le laboratoire MIND de l'université du Maryland [Kalyanpur, 06]. Il est comme Protégé 2000, développé en Java mais il est moins complet et non plus extensible. Le SWOOP sert à visualiser presque toutes les ontologies qui ne peuvent pas ouvrir avec Protégé. Il est basé sur la représentation sous forme arborescente des classes et de leurs propriétés. Le contenu d'un élément sélectionné est représenté dans un frame séparé cela nous a permis de connaître rapidement les instances d'une classe [Kalyanpur, 05]. De plus, il permet de faire des statistiques sur une ontologie donnée, à savoir, le nombre de classes, le nombre de propriétés, le nombre des instances, le niveau de granularité maximal, etc. [Kalyanpur, 05].
- ***OntoEdit* :** OntoEdit (Ontology Editor) est un éditeur extensible et payant dans sa version complète, ses ontologies sont représentables sous forme d'un arbre hiérarchique [Sure, 02]. Il dispose un gestionnaire des questions de compétences via l'outil ontokick pour lesquels l'ontologie doit fournir des réponses [Sure, 02]. OntoEdit intègre un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs avec la prise en compte d'un contrôle de cohérence et la gestion des ordres d'édition [Sure, 03].

De nombreux outils permettent aujourd'hui de construire des ontologies. Dans notre travail nous utiliserons l'éditeur Protégé 2000 qui représente l'outil le plus utilisé dans nos jours. De plus, il offre une interface utile et utilisable permettant de manipuler aisément tous les éléments d'une ontologie OWL et la disponibilité des plugins pour assurer une gestion efficace des ontologies. Le Protégé 2000 est également une librairie Java pour créer des applications à bases de connaissances via les API de manipulation des ontologies, tels que Jena, Sesame et Corese, etc. [Noy, 00a]. Il permet aussi de gérer les données multimédias en associant des annotations sémantiques de ce type de donnée. L'annotation peut être effectuée par deux manières différentes [Noy, 00a] : une représentation textuelle du contenu en utilisant l'outil Image Widget ou bien une représentation formelle exprimée à l'aide des concepts, des relations et des instances décrits dans une ontologie.

10.2. Outils d'alignement d'ontologies

De nombreux outils et Framework permettent d'aligner les ontologies à titre d'exemple : PROMPT, OLA, FOAM, ASMOV, WeSeE, GOMMA, HotMatch, Wikimatch, Optima, etc. Dans cette thèse, nous représentons les trois premiers outils qui sont les plus utilisés dans le domaine du web sémantique:

- ***PROMPT* :** Noy et Musen [Noy, 00b] ont proposé un algorithme et un outil graphique appelé Anchor-PROMPT ou tout simplement PROMPT pour l'automatisation de fusion et d'alignement des ontologies. L'outil PROMPT considère l'ontologie à aligner comme

un graphe étiqueté orienté où ses nœuds sont des classes d'ontologie et les arcs représentent les relations entre les classes. L'alignement d'ontologie est basé sur l'utilisation d'une liste des classes similaires 'matchers' définies manuellement par des utilisateurs ou automatiquement par la mise en correspondance lexicologique [Noy, 00b]. L'algorithme d'alignement Anchor-PROMPT analyse les chemins des graphes dans un premier temps puis il détermine les classes les plus fréquemment survenues dans les mêmes positions sur les chemins similaires [Noy, 00b].

- ***OLA:*** l'outil OLA (OWL Lite Alignment) sert à aligner automatiquement les ontologies décrites en OWL-Lite et OWL-DL [Euzenat, 04]. Son principe consiste à appliquer des mesures de similarité partielles entre les éléments (classes, propriétés, relation, etc.) de deux ontologies. La similarité globale est la somme de toutes les similarités partielles calculées. De plus, OLA est muni d'un algorithme d'optimisation de similarité partielle qui s'appuie sur le calcul itératif du point fixe [Euzenat, 04].
- ***FOAM*:** le FOAM (Framework for Ontology Alignment and Mapping) est un Framework dédié non seulement à l'alignement des ontologies mais aussi de faire des mappings entre ontologies [Ehring, 07]. Ce Framework est comme l'outil OLA, basé sur le calcul de similarité globale à partir des similarités partielles de chaque paire d'entités des ontologies à aligner. De plus, il offre à l'utilisateur la possibilité d'accepter ou de rejeter des recommandations d'alignement. Ce Framework est très utile dans des applications d'intégration de données, d'évolution d'ontologies, de fusion d'ontologies, etc.

Ces outils d'alignement diffèrent au niveau de la stratégie d'alignement, des mesures de similarité entre deux entités ainsi que de la combinaison de ces mesures. Dans le contexte de notre travail, nous n'avons pas besoin d'utiliser un tel outil d'alignement pour assurer l'intégration de données. Dans cette thèse, nous désirons de proposer un nouveau processus d'alignement d'ontologie qui doit être utilisé dans un système de médiation sémantique dédié au traitement de l'hétérogénéité sémantique aux bases de données multimédias et hétérogènes.

11. Conclusion

Les ontologies ont pour but de représenter les connaissances d'une partie du monde réel à partir d'une spécification explicite et formelle d'une conceptualisation. Quel que soit le type d'ontologie à développer, plusieurs méthodologies ont été proposées pour la construction d'ontologies. Ces méthodologies reposent généralement sur deux étapes fondamentales: la définition de la hiérarchie des concepts composant l'ontologie et l'opérationnalisation de cette ontologie. Dans la première étape, trois approches ont été proposées pour hiérarchiser les concepts d'ontologie : approches ascendante, descendante et intermédiaire. Dans la seconde étape, plusieurs langages de représentation et de manipulation d'ontologie, les plus importants sont OWL et SPARQL. De plus, PROTEGE 2000 est l'un des outils les plus populaires pour l'édition et la gestion des ontologies.

Dans le cadre de notre travail, nous appuierons sur l'approche de Noy et Mcguinness pour la construction de nos ontologies à partir de zéro. Cette approche est basée sur un

processus itératif tout au long de cycle de vie d'ontologie. Ainsi, nous utiliserons l'approche descendante (*top-down*) pour la définition de la hiérarchie des concepts de nos ontologies de domaine.

Par ailleurs, l'utilisation de l'ontologie lexicale WordNet est très importante pour définir les relations lexicales entre termes. Elle est souvent utilisée conjointement avec une ontologie pour désambiguïser le sens des mots et aligner deux ontologies ayant besoin de communiquer et partager les connaissances. Plusieurs techniques et outils d'alignement d'ontologies ont été proposés et qui sont basés sur l'utilisation d'une mesure de similarité sémantique entre les entités de deux ontologies.

L'alignement d'ontologie s'impose comme une solution, afin de permettre l'interopérabilité et le partage des sources de données hétérogènes. Dans ce contexte, nous allons présenter dans la deuxième partie de ce manuscrit (partie contributions), notre processus d'alignement d'ontologie qui est utilisé dans notre système de médiation sémantique pour le traitement du problème d'hétérogénéité sémantique des bases de données multimédias et hétérogènes.

PARTIE II.

CONTRIBUTIONS

Cette partie aborde nos contributions et expérimentations effectuées. Elle est composée de deux chapitres ; le premier est consacré à présenter la première approche proposée avec ses expérimentations. Le deuxième chapitre expose la deuxième approche dédiée à l'indexation personnalisée de documents multimédias et scientifiques. Il présente également une étude expérimentale et une analyse des différents résultats obtenus.

CHAPITRE

04

APPROCHE PROPOSÉE POUR LA MÉDIATION SÉMANTIQUE DES BDMM: PRÉSENTATION ET EXPÉRIMENTATIONS

Ce chapitre présente la première approche proposée dans cette thèse dont le but du traitement d'hétérogénéité sémantique pour la médiation sémantique des bases de données multimédias.

1. Introduction

Ce chapitre présente notre première contribution concernant le traitement d'hétérogénéité sémantique pour l'exploration des sources de données multimédias. Nous présentons une nouvelle approche du traitement de requête dans un système de médiation sémantique dédié à l'intégration des bases de données multimédias et hétérogènes (BDMM). Le problème d'hétérogénéité sémantique de ce type de sources de données se situe aux deux niveaux : hétérogénéité niveau requête et hétérogénéité niveau source de données.

Une deuxième partie de ce chapitre présente une étude expérimentale pour valider, évaluer et analyser cette première proposition. De ce fait, nous allons présenter dans un premier lieu, l'implémentation du système de médiation sémantique ***SAMER*** (**S**em**A**ntic **M**ediation for alim**E**ntation **R**isks) en utilisant le langage de programmation JAVA sous l'environnement de développement Eclipse Luna et nous montrons le déroulement de requêtes d'utilisateur. Dans un second lieu, nous allons mener d'une évaluation des performances en termes de qualité de données intégrées et nous discuterons les résultats obtenus. Cette étude expérimentale se termine par l'analyse de résultats de deux types de comparaison : une comparaison quantitative qui porte sur l'importance des mesures de similarité utilisées dans notre approche et une comparaison qualitative qui vise à comparer notre travail avec d'autres travaux similaires.

2. Aperçu général de l'approche proposée

Notre approche permet d'explorer aisément des bases de données multimédias et hétérogènes (BDMM) en traitant le problème d'hétérogénéité sémantique selon deux niveaux : hétérogénéité niveau requête et hétérogénéité niveau sources de données. L'approche proposée s'appuie sur le principe de l'approche de médiation sémantique par hybridation et l'approche LAV (Local As View) pour définir le mapping entre le schéma global et les schémas locaux des sources de données [Aggoune, 17]. Partant de ce fait, notre approche repose sur la construction et l'utilisation d'ontologie de domaine de risque alimentaire ONTARIS (ONTology of Alimentation RISks) en tant qu'ontologie partagée de la couche médiateur et les ontologies virtuelles de chaque BDMM qui ont été construites automatiquement à partir des BDMM. Ces ontologies virtuelles jouent le rôle des vues sémantiques permettant d'assurer l'intégration sémantique de données. De plus, l'ontologie partagée ONTARIS est le schéma global du médiateur qui fournit un accès unifié aux sources de données via le langage SPARQL. Elle permet aussi d'enrichir ces différentes BDMM par des concepts importants décrivant le domaine sur lequel les BDMM sont construites.

Avant de présenter en détail l'approche proposée, nous devons décrire dans un premier temps le domaine d'application pour construire nos bases de données multimédias qui sont modélisées par à la fois le modèle relationnel et le modèle orienté-objet. Dans un second temps, nous allons présenter notre algorithme de construction automatique des ontologies virtuelles à partir d'une base de données, puis nous procéderons à la description des différentes étapes de construction de l'ontologie partagée ONTARIS dédiée au domaine de risques alimentaires que nous comptons utiliser comme schéma global de notre système de médiation sémantique.

3. Description du domaine des risques alimentaires

Afin de valider notre contribution pour le traitement du problème d'hétérogénéité sémantique des sources de données hétérogènes, il est nécessaire de spécifier le domaine d'application sur lequel on crée des sources de données multimédias et hétérogènes. En effet, nous avons choisi l'un des domaines de la bioinformatique, il s'agit le domaine d'étude des risques microbiologiques dans les aliments ou tout simplement les risques alimentaires.

Dans ce cadre, nous présentons en premier lieu les motivations de choix de ce domaine et dans un second lieu, nous donnons sa description.

3.1.Motivations

Le domaine des risques alimentaires est un domaine en plein essor, qui consiste à analyser et rechercher les risques alimentaires qui peuvent toucher la santé humaine de la population [Feinberg, 06]. Il peut référencer d'autres domaines à titre d'exemple, la toxicologie, les sciences des aliments et de nutrition, l'économie agroalimentaire qui étudie les procédés de transformation et de conservation des aliments, les sciences de la consommation, l'épidémiologie qui traite les pathogènes et les facteurs influant sur la santé, la médecine générale et les maladies résultantes de la consommation d'aliments contaminés avec les traitements possibles [Feinberg, 06]. Le choix du domaine de risques alimentaires est motivé par les cinq raisons suivantes :

- ***La richesse d'informations*** : de nombreuses informations peuvent être utilisées dans le domaine des risques alimentaires à savoir [de Valk, 15], les aliments (plus de 50000 types), les microorganismes (plus de 2 millions), les facteurs influant sur la propagation des microorganismes, des centaines de maladies infectieuses et toxiques ainsi que les antibiotiques pour traiter les infections provoquées par des microorganismes.
- ***La diversité des sources de données*** : toutes ces informations sont apparues dans différentes sources de données qui sont souvent hétérogènes (bases de données, sites web, des ontologies, des fichiers Excel, des tableaux, etc.). Cette diversité des sources peut provoquer des problèmes d'hétérogénéités sémantiques lors de l'exploration de ces données.
- ***La disponibilité de type de données multimédias*** : les données de domaine de risques alimentaires peuvent être de données textuelles (nom de produits alimentaires, nom d'un virus, etc.), de données images (les images des microorganismes, les aliments, etc.), des séquences vidéo décrivant ces informations avec des extraits sonores informant les risques qu'ils peuvent survenir dans des aliments contaminés.
- ***La possibilité de faire des préventions :*** effectuer des analyses sur la prolifération des microorganismes dans les produits alimentaires, élaborer une évaluation empirique des risques et établir des statistiques sur le taux d'apparition des microorganismes et leurs facteurs qui favorisent leur coexistence, etc. [Feinberg, 06]. Toutes ces opérations nous ont permis de faire des préventions sur les microorganismes qui vont se multiplier, survivre ou mourir dans l'aliment. Alors, mieux vaut prévenir que guérir les maladies dues aux aliments.

- ***La sécurité alimentaire*** : La connaissance et l'identification des risques et des dangers alimentaires qui peuvent menacer la santé publique, nous ont permis d'assurer une meilleure sécurité possible des produits alimentaires. L'amélioration de la sécurité alimentaire est donc essentielle pour atteindre les objectifs de développement durable.

Par conséquent, pour réaliser notre travail nous construisons et nous utilisons les bases de données multimédias et hétérogènes issues de domaine des risques alimentaires.

3.2.Description

Consommer des aliments salubres et nutritifs est essentiel pour vivre en bonne santé. Toutefois, les mauvaises utilisations et conservations des aliments ont conduit à l'apparition des microorganismes qui risquent de porter atteinte à notre santé. Prévoir les niveaux de risque dans les aliments permet de garantir la salubrité des aliments [Feinberg, 06].

De plus, consommer des produits alimentaires (viandes, végétales, boissons, etc.) contaminés provoque plus de 200 maladies, allant de la diarrhée au cancer [Feinberg, 06]. Ces maladies sont généralement infectieuses ou toxiques par nature et provoquées par des agents pathogènes tels que les bactéries, les virus, les parasites ou des substances chimiques.

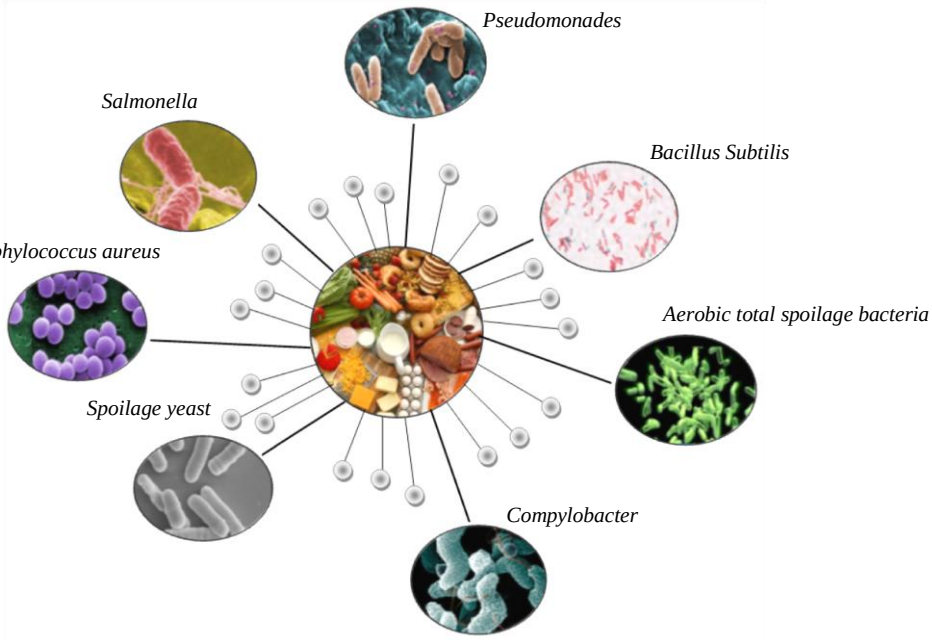


Figure 4.1. Illustration graphique de domaine de risques alimentaires

L'organisation mondiale de la santé OMS[15] a donné des statistiques annuelles en l'an 2015 et elle estime que plus de 600 millions de personnes sont tombées malades à cause de la consommation des aliments contaminés, qu'il y a eu près de 420 000 décès. Les principales maladies d'origine alimentaire sont [de Valk, 15]: la diarrhée et la diarrhée grave, les

[15] http://www.who.int

infections, la méningite, l'empoisonnement grave et les maladies à long terme comme le cancer.

Prenons l'exemple de salmonellose ; une maladie infectieuse qui provoque des troubles digestifs à cause de l'existence des bactéries appelées salmonelles (Salmonella) [Marano, 13]. Ces bactéries infectent le tube digestif de l'être humain qui consomme des aliments contaminés d'origine animale (viandes, œuf ou lait) qui peuvent être crus ou peu cuits. Ils peuvent rarement apparaître dans les fruits frais ou les légumes crus contaminés.

Par ailleurs, le développement récent de l'analyse des risques alimentaires et la définition des normes et des référentiels ont fait émerger un besoin des systèmes d'information permettant un accès unifié à ses différentes bases scientifiques, avec l'objectif de connaître et prévoir les atteintes possibles à la santé publique [de Valk, 15].

Dans ce contexte, nous présenterons dans cette thèse notre système de médiation sémantique des sources de données multimédias et hétérogènes. Il est donc nécessaire de définir les différentes sources de données utilisées ainsi que l'ontologie ONTARIS (ONTology of Alimentation RISks) de domaine de risques alimentaires.

4. Représentation des BDMM et la construction des ontologies du médiateur

Dans le but de valider et d'évaluer notre système de médiation sémantique, nous avons créé six bases de données multimédias (BDMM) ; quatre sont modélisées par le modèle relationnel et deux autres sont modélisées par le modèle orienté-objet. Chaque base de données a été construite par un concepteur des bases de données (BD). De ce fait, six concepteurs modélisent les BDMM de manières différentes et ne produiront pas systématiquement la même base de données. En effet, ces bases de données sont hétérogènes selon différents aspects (structure, syntaxe, sémantique, etc.).

Le choix des BDMM comme mode de représentation de données multimédias est motivé par le fait qu'elles sont des supports puissants pour le stockage, la modélisation et l'exploration des données. Grâce aux systèmes de gestion de bases de données (SGBD) on peut traiter de gros volumes de données sous une forme structurée. De plus, ils disposent d'un langage de requêtes normalisé permettant d'exprimer de manière déclarative des interrogations à la base.

Nous présentons dans ce qui suit ces bases de données qui seront localisées au niveau de la couche de sources de données de notre système de médiation sémantique.

4.1. Les BDMM à base du modèle relationnel

Nous avons créé quatre bases de données multimédias qui sont fondées sur le modèle relationnel qui a été inventé par l'informaticien britannique *Edgar Frank Codd* en 1970 [Codd, 70].

Le modèle relationnel consiste à représenter les données sous forme des tables ou relations reliées entre elles et ayant un nom et un ensemble des attributs avec leurs types et des contraintes d'intégrité (clé primaire, clé étrangère et la condition check) [Codd, 70]. Les

lignes de ces relations sont appelées des n-uplets (ou tuples en anglais) ou des enregistrements.

Pour créer ces bases de données, nous avons utilisé le SGBD Oracle Database[16] 10g Express Edition pour ses performances, sa fiabilité et sa sécurité.

Oracle Database 10g a été conçue pour la technologie Grid Computing qui permet de regrouper les serveurs de sorte qu'ils ne forment qu'une seule entité plus puissante. Il s'appuie sur une architecture à trois niveaux: le poste client pour gérer la représentation des données via un navigateur Web, le serveur d'applications qui gère la logique d'application et le serveur de données vise à gérer le stockage des données et les transactions orientées données [Delmal, 00].

Les données d'une base de données d'Oracle sont stockées dans un tablespace composé d'un dictionnaire de données qui représente le cœur de la base de données [Delmal, 00].

Nous utilisons également le type BLOB pour définir le domaine du média image et nous créons le répertoire ImgAlimRisks pour arranger les fichiers images de toutes les BDMM.

La figure suivante représente d'une part, les tables relationnelles de l'une des BDMM utilisées dans ce travail et d'autre part, un exemple d'une commande SQL* Plus d'Oracle pour créer la table relationnelle Food.

```
Entrez le nom utilisateur : system
Entrez le mot de passe :
Connectú.
SQL> CREATE OR REPLACE DIRECTORY ImgAlimRisks AS 'd:/imagesAR';

Rúpertoire crúú.

SQL> CREATE TABLE Food (NumRO integer Primary key,
  2  NameFO VARCHAR(60),
  3  TypeFO VARCHAR(60),
  4  Species VARCHAR(60),
  5  Familly VARCHAR(60),
  6  OrcderFO VARCHAR(30),
  7  BinominalNameFO VARCHAR(100),
  8  Calorie INTEGER,
  9  DateofDiscovery DATE,
 10  Minerals VARCHAR(200),
 11  Vitamins VARCHAR(200),
 12  PhotoFO BLOB);

Table crúúe.
```

Commande SQL*PLUS

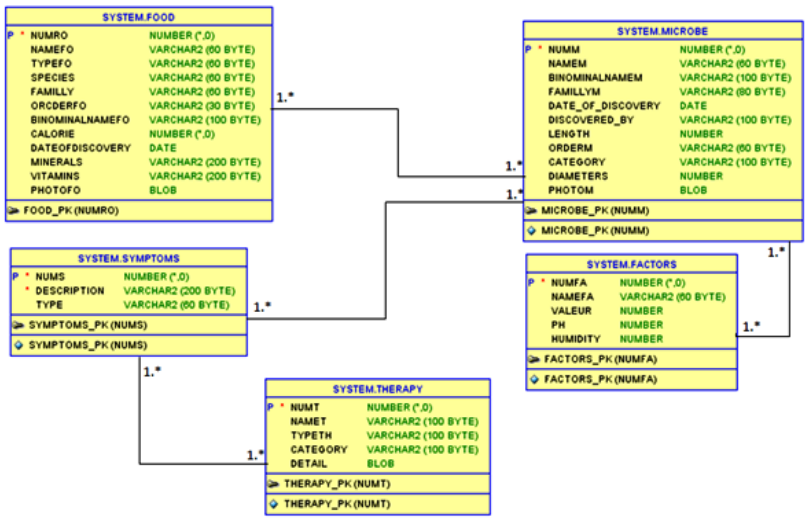


Figure 4.2. Les tables relationnelles de l'une des BDMM utilisées

Toutes ces bases de données sont hétérogènes et donc plusieurs problèmes d'hétérogénéité sémantiques peuvent être survenus. Ces problèmes ont été résumés dans la table suivante:

[16] www.oracle.com

Problèmes d'hétérogénéité sémantique	Exemples
Diversité de structures de données	• Pour représenter les données du domaine de risques alimentaires, l'une des bases de données est composée de cinq tables tandis qu'une autre est seulement par deux tables; • Le nombre et le type des attributs contenant la table Food (par exemple) sont différents dans les quatre bases de données; • Dans la table Factor, l'attribut PH est associé à une contrainte check sur sa valeur (ex. CHECK (PH BETEWEN 1 AND 14)) par contre dans une autre base contenant cet attribut ne l'est pas; • La table Microbe possède une méthode NB qui calcule le nombre des microbes qui peuvent être existés dans un aliment, tandis que la même table dans une autre base ne contient pas cette méthode.
Diversité de définitions d'un même type d'entité	• Problème de synonymes de nom des tables (ex. Food et Nutriment); • Problème de synonymes de nom des attributs (ex. kind et type); • Problème de redondance entre un mot et son acronyme (ex. Veggietbale et Veggie).
Les erreurs syntaxiques	• Un terme désignant une entité peut être syntaxiquement erroné ou incomplet (ex. Vitamin et Vitam).
Diversité des médias utilisés pour un même type d'attribut	• Dans ces quatre bases de données, l'attribut détail de la table Symptom peut être représenté par un type CLOB ou par une image illustrative (BLOB).
Mise à l'échelle de données	• L'attribut température de la table *Factor* peut être donné en degré Celsius tandis qu'il est défini en degré *Fahrenheit* dans une autre base de données.
Pauvreté sémantique du modèle relationnel	• Le modèle relationnel est fondé sur la théorie mathématique des ensembles cela implique qu'il est pauvre en capacité de représentation sémantique.

Table 4.1. Problèmes d'hétérogénéité sémantique dans les BDMM basées sur le modèle relationnel

Tous ces problèmes impliquent pratiquement une hétérogénéité sémantique lors de l'intégration des BDMM dans le système de médiation. De plus, ces problèmes apparaissent entre les sources de données (bases de données), ils peuvent se retrouver entre les ontologies qui jouent le rôle d'une ressource sémantique. Néanmoins, l'intégration de données à base d'ontologies permet d'assurer l'interopérabilité qui devient plus souple à gérer avec le modèle ontologique qu'avec le modèle relationnel ou orienté-objet.

4.2. Les BDMM à base du modèle orienté-objet

Deux concepteurs de la base de données ont créé chacun une BDMM à base du modèle orienté-objet. Ces bases de données regroupent sous forme de classes d'objets persistants. Ces objets ayant une identité OID (Object IDentifier) qui le distingue de tout autre objet, même s'ils ont la même valeur. La création de ces deux bases de données est faite par le SGBD orienté-objet O2 basé sur le langage C [Bancilhon, 88] de la société O2 Technology qui permet la définition de classes avec leurs attributs, méthodes et constructeurs de type tuple, list, set et class. Les classes et les méthodes sont stockées dans des schémas. O2 possède de nombreuses classes prédéfinies qui peuvent être utilisées pour la réalisation de schémas, les plus importantes sont : la classe *Image* avec sa méthode loadfile("nomfichier.gif"), la classe

Bitmap pour charger les images bitmap (*.BMP) et la classe *Text* pour la définition de données textuelles [Bancilhon, 88]. Il permet aussi de gérer l'héritage multiple entre classes.

La définition des classes se fait par le langage de définition d'objet ODL (Objet Definition Language) et l'interrogation de BDOO se fait à l'aide de langage de requêtes proche de SQL appelé OQL (Object Query Language) [Amann, 93]. La figure suivante présente le schéma objet de l'une des BDMM à base du modèle orienté-objet.

```
Schema Alimentation-Risks;
CLASS Aliment
EXTENT Aliments {NumA: integer,
NameA: string, Type: string,
Binominal-name: string, specie: string,
FamillyA: string, order: string, Country: string,
Calorie: integer, data-discovery: Date,
Minerals: SET (string), Vitamins: SET (string),
Photo: Image, RELATIONSHIP SET Germ Has-germ
INVERSE Germ:: Has-aliment} end;
CLASS Germ
EXTENT Germs {NumG: integer,
NameG: string, TypeG: string,
Binominal-name: string, category: string,
FamillyA: string, order: string,
Discovered-by: string, data-discovery: Date,
Length: real, Diameter: real,
Photo: Image, RELATIONSHIP SET Aliment
Has-aliment INVERSE Aliment:: Has-germ,
RELATIONSHIP SET Factor
Has-factor INVERSE Factor:: Has-germ-Fact,
RELATIONSHIP SET Indice Has-indice INVERSE
Indice:: Has-germ-indice
} end;

CLASS Factor
EXTENT Factors {NumFa: integer,
NameFa: string, Value: real,
PH: integer, humidity: real,
RELATIONSHIP SET Germ Has- germ-Fact
INVERSE Germ:: Has-Factor} end;
CLASS Indice
EXTENT Indices {NumS: integer,
NameS: string, Type: string,
RELATIONSHIP SET Germ Has-germ-indice
INVERSE Germ:: Has-germ,
RELATIONSHIP SET Treatment
Has-treatment INVERSE Treatment::
Has-Tre-indice} end;
CLASS Treatment
EXTENT Treatments {NumT: integer,
NameT: string, TypeT: string,
Category: string, Detail: Image,
RELATIONSHIP SET Indice Has-Tre-indice
INVERSE Indice:: Has-treatment} end;
```

Figure 4.3. Les classes d'une BDMM à base du modèle orienté-objet

Les objets de la classe "Aliment" se trouvent dans la collection "*Aliments"***.** Grâce à cette collection, on peut interroger la base de données orientée-objet, par exemple, pour afficher tous les aliments existants on écrit tout simplement le nom de la collection "Aliments" au lieu de faire une requête SELECT, afficher le nom de Treatment on exécute la requête OQL suivante : SELECT T.NameT FROM T in Treatments.

Il est clair que ces BDMM sont hétérogènes et tous les problèmes d'hétérogénéité sémantiques survenus dans le modèle relationnel (cf. Table 4.1) sont aussi apparus dans les BDMM à base du modèle orienté-objet.

Utiliser comme sources de données à intégrer des bases de données multimédias modélisées par le modèle relationnel et celles modélisées par le modèle orienté-objet engendrent trois problèmes importants : la diversité du modèle de données (relationnel et objet), la diversité du SGBD utilisé (ORACLE et O2) et les différents langages de définition et de manipulation de données (SQL pour le modèle relationnel, et ODL et OQL pour le modèle orienté-objet).

Pour traiter ces différents problèmes d'hétérogénéité sémantique, nous associons à chacune de ces bases de données sa propre ontologie locale contenant leurs connaissances.

4.3. Construction et manipulation des ontologies virtuelles à partir des BDMM

Les six bases de données que nous avons créées sont reliées avec sa propre base de connaissances qui est représentée par une ontologie virtuelle. Cette dernière est vue comme une ontologie locale qui fournit des vues sémantiques (*semantic views*) à sa base de données. L'objectif visé est d'offrir des vues sémantiques écrites en même langage OWL que le schéma global du médiateur (l'ontologie partagée ONTARIS) afin de faciliter l'intégration sémantique de données hétérogènes.

Les bases de données à base d'ontologie ont été définies dans divers travaux comme ceux de Bellatreche et al [Bellatreche, 03], Pierra et al [Pierra, 05], Krivine et al [Krivine, 09] et Sultan et al [Sultan, 13]. Ces travaux visent à définir une base de données à base d'ontologie comme une base qui pouvant stocker à la fois le contenu usuel d'une base (les données) et l'ontologie qui décrit la signification de ces données. La majorité de ces travaux sont basés sur l'utilisation des outils existants dédiés à la transition entre bases de données et ontologies comme: DataMaster[17] de protégé, KAON2[18] et RDBToOnto[19]. Ces outils permettant de créer une ontologie par la transformation des tables d'une base de données en des concepts OWL et chaque n-uplet ou enregistrement de la table devient un individu du concept correspondant [Krivine, 09]. Néanmoins, ces outils nécessitent un espace mémoire important pour stocker l'ontologie obtenue et ils ne permettent pas de gérer l'évolution de ces ontologies ; une fois l'ontologie a été construite on ne peut pas la mettre à jour au fur et à mesure de sa base de données.

Dans notre travail, nous proposons un algorithme de construction automatique des ontologies virtuelles à partir d'une base de données [Aggoune, 17]. Ces ontologies sont utilisées comme des vues sémantiques qui décrivent le contenu des bases de données. Nous utilisons également une colonne virtuelle (ou attribut virtuel) dans chaque table d'une BDMM afin de gérer l'évolution et la cohérence du contenu de la BDMM avec celui de son ontologie virtuelle.

Cet algorithme prend en entrée la base de données relationnelle (ou orienté-objet) et donne en sortie une ontologie virtuelle. Prenons l'exemple d'une base de données relationnelle : pour chaque table relationnelle C[i], on va extraire les informations suivantes :

- Ses attributs (ou nom de colonne) par la requête SELECT COLUMN_NAME FROM USER_TAB_COLUMNS WHERE TABLE_NAME = C[i]. Ces noms de colonnes ont été sauvegardés dans la table P [i, j] où i et j indiquent respectivement, le numéro de la table et le numéro de l'attribut;
- Les types T [i, j] de chaque colonne pour convertir ces types aux types propres de ProtégéOWL en utilisant l'algorithme 4.2;
- Ses n_uplets V [i, k] qui deviennent des individus du concept i.

[17] http://protegewiki.stanford.edu/index.php/DataMaster
[18] http://kaon2.semanticweb.org/
[19] http://www.tao-project.eu/researchanddevelopment/demosanddownloads/RDBToOnto.html

```
Algorithme construction d'ontologie virtuelle
Entrées : BD : base de données ;
          VO : ontologie virtuelle initialisée à vide
Début
  C :=une table du nom des relations de BD;
  Pour chaque élément C[i] faire
    VO. concept := Extraire-Table(C[i]) ;
    P [i, j] := Extraire-Attribut (C[i]) ;
    T [i, j] := Extraire-Type (P [i, j]) ;
    V [i, k] :=Extraire-données (C[i]) ;
    Pour chaque élément P [i, j] faire
      Ajouter P [i, j] comme propriétés de VO;
      Ajouter_type (T [i, j]) dans VO;
    Fin pour
    Pour chaque élément V [i, k] faire
      Ajouter V [i, k] comme instances de VO;
      Créer une colonne virtuelle CV dans la i ième table relationnelle;
      Insérer dans CV URI de l'instance de VO;
    Fin pour
  Fin pour
  Vider (C, P, V, T) ;
Fin
Sortie : VO : ontologie virtuelle
```

Algorithme 4.1. Algorithme de construction automatique d'ontologie virtuelle

En ce qui concerne la conversion de type d'attribut, l'algorithme de construction automatique d'ontologie virtuelle fait appel à notre procédure Ajouter_type (T [i, j]) qui prend comme paramètre le tableau des types d'attribut du rang j de la $i^{ème}$ relation. Cette procédure permet de gérer tous les types existants (integer, varchar, string, blob, image, etc.).

```
Algorithme Ajouter_type (T : tableau de n lignes et m colonnes)
Entrées : VO : ontologie virtuelle
Début
  Pour chaque élément T [i, j] faire
    SI T [i, j]='Varchar' ou 'Char' ALORS
       Type de propriété dans VO est STRING
    SINON SI  T [i, j]='Blob' ou T [i, j]='Image' ou T [i, j]='Clob'  ALORS
      Début
        String chemin := récupérer le chemin de l'image;
        Type de propriété dans VO est chemin;
      Fin sinon si
    Sinon Type de propriété dans VO est T [i, j] ;
  Fin pours
Fin.
```

Algorithme 4.2. Algorithme de gestion de type d'attributs

Dans cet algorithme, les attributs du type complexe tels que BLOB, Image sont devenu de type string dont sa valeur est le chemin où se trouve le fichier image dans le répertoire ImgAlimRisks.

En revanche, pour assurer la cohérence entre le contenu de base de données et celui de son ontologie virtuelle VO, nous utilisons une colonne virtuelle CV dans chaque table relationnelle. Cette colonne est vue comme un pointeur vers l'ontologie afin de lier chaque n-uplet de la BD par son instance dans l'ontologie virtuelle (cf. algorithme 4.1.). Le contenu de cette colonne est donc l'URI (Uniform Resource Identifier) d'instance dans VO. À partir de

cette colonne virtuelle on peut effectuer sur l'ontologie virtuelle par le langage SPARQL toutes les opérations qui ont été appliquées à sa base de données. L'algorithme suivant, illustre l'évolution de l'ontologie par l'évolution de sa base de données.

```
Algorithme Evolution VO
Entrées : BD : base de données, VO: ontologie virtuelle
          Op : {'aucun', 'ajout', 'suppression', 'mise à jour'}
Début
  Pour chaque i^ème table de la BD FAIRE
  Début
    SI op= 'aucun' Alors Exit
    SINON SI op= 'Ajout' Alors
      Début
          Ajouter dans VO la Nouvelle instance;
          Ajouter dans CV URI de nouvelle instance ;
       Fin si
    Sinon si op= 'suppression' Alors
      Début
          URI := CV[k]    //extraire l'uri de K^ème élément à supprimer
          Supprimer de VO l'instance de URI;
          Supprimer CV[k];
       Fin si
    SINON     //mise à jour
      Mettre à jour dans VO l'Instance k;
   Fin pour
Fin.
```

Algorithme 4.3. Algorithme d'évolution d'ontologie virtuelle

Les trois algorithmes que nous avons proposés pour la construction et la gestion d'ontologies virtuelles à partir d'une base de données, sont utilisés dans notre approche de traitement d'hétérogénéité sémantique de données multimédias et hétérogènes.

Construire une ontologie virtuelle plutôt d'importer la base de données en ontologie et les utiliser séparément, permet de réduire l'espace mémoire dans le système de médiation sémantique par rapport les systèmes existants. De plus, ces ontologies virtuelles jouent le rôle des vues abstraites décrivant la sémantique du contenu des bases de données et qui sont exprimées par le même langage OWL du schéma global. Ces vues sémantiques visent à réduire à la fois la charge de travail au niveau de la couche médiateur et la couche adaptateurs.

Dans la couche médiateur, on n'a pas besoin de réécrire la requête initiale écrite en SPARQL selon le vocabulaire du schéma global en termes des vues sémantiques car, l'exécution de requête dans notre système revient à faire des appariements entre les composants de requête et les entités des ontologies virtuelles (vues sémantiques). Une fois la requête initiale a été soumise au médiateur, elle a été transmise directement aux adaptateurs.

Dans la couche adaptateurs, on n'a pas besoin de traduire les requêtes venues de la couche médiateur en des requêtes exprimées en termes du vocabulaire de langage de bases de données (SQL pour la BD relationnelle et OQL pour la BD orienté-objet). Chaque adaptateur doit appliquer notre processus d'appariement du contenu des requêtes et celui des ontologies virtuelles (référencer le lecteur à la section 6.2).

Par ailleurs, utiliser une colonne virtuelle dans chaque table relationnelle (ou attribut virtuel dans chaque classe de BD orienté-objet) permet d'assurer la dépendance entre chaque base de données et son ontologie virtuelle. De ce fait, toutes modifications apportées sur la base de données seront appliquées sur l'ontologie virtuelle correspondante.

Les avantages de ces propositions nous ont permis de réduire considérablement le temps d'exécution d'une requête dans notre système de médiation sémantique. La prochaine sous section présente la construction de notre ontologie partagée ONTARIS qui est utilisée comme schéma global du système de médiation sémantique.

4.4. Construction de l'ontologie ONTARIS

Notre système de médiation sémantique suit le principe de l'approche de médiation sémantique par hybridation. Cette approche vise à utiliser une ontologie principale dite ontologie partagée qui représente le schéma global du système de médiation et des ontologies locales associées à chaque source de données décrivant leur sémantique (cf. chapitre 2, section 7.3.). Dans ce contexte et après avoir présenté dans la section précédente comment construire automatiquement une ontologie virtuelle à partir d'une base de données, nous décrivons dans cette section, les étapes de construction de l'ontologie partagée ONTARIS.

Comme nous avons mentionné dans le chapitre précédent que la construction de notre ontologie de domaine est faite en partant du zéro, suivant l'approche de Noy et Mcguinness et nous utilisons l'approche top-down pour la définition de la hiérarchie de ses classes.

L'ontologie de domaine de risques alimentaires, baptisée **ONTARIS** (**ONT**ology of **A**limentation **RIS**ks) est l'ontologie partagée de notre système de médiation sémantique qui offre à l'utilisateur un support d'exploration unifiée et sémantique via le langage standard SPARQL.

L'adaptation du processus de Noy et Mcguinness à notre travail, nous a permis de définir le processus résumé en quatre principales étapes [Aggoune, 17]:

1. *Spécification des besoins* : après avoir choisi le domaine qui va couvrir l'ontologie, nous énumérons les différents termes, concepts, propriétés et relations composant l'ontologie, et s'il possible de réutiliser des ressources existantes (ontologie, thésaurus, WordNet, etc.);
2. *Conceptualisation* : à partir de l'approche top-down, nous définissons les classes, la hiérarchie des classes et les instances afin de créer le modèle conceptuel d'ontologie;
3. *Encodage* : nous ajoutons deux autres étapes : l'encodage et l'évaluation. L'encodage consiste à représenter les connaissances obtenues dans l'étape précédente en langage formel OWL en utilisant l'éditeur Protégé 2000;
4. *Evaluation* : permet de tester l'adéquation de l'ontologie, ainsi que la maintenance en cas de modification de composants.

4.4.1. Spécification des besoins

Cette étape regroupe les trois premières étapes de l'approche de Noy et Mcguinness. En effet, dans cette étape on identifie les termes les plus importants de domaine de risques alimentaires tels que : aliments, microorganismes, les facteurs d'existence de ces microorganismes dans les aliments, les symptômes d'intoxication et les maladies infectieuses lors de consommation d'un aliment contaminé avec quelques traitements nécessaires pour guérir ces maladies.

La collecte de ces termes est faite à partir d'utilisation de documentation en ligne et d'effectuer des interviews avec des biologistes, des microbiologistes, des spécialistes dans le domaine de la toxicologie et éventuellement des médecins.

Nous utilisons aussi l'ontologie lexicale WordNet comme ressources externes pour éviter toutes redondances entre les termes à cause de relation de synonymies entre eux.

Le résultat de cette étape est un dictionnaire des termes avec leur description. La table suivante représente deux éléments de ce dictionnaire.

Terme	Description
Microbe	Décrit les microbes vivants dans un aliment selon un certain nombre des facteurs.
Therapy	Décrit le traitement à prendre lorsqu'une personne est infectée par un aliment contaminé.

Table 4.2. Un fragment du dictionnaire de termes

4.4.2. Conceptualisation

Suivant l'approche top-down pour la définition de la hiérarchie des concepts d'ONTARIS, nous commençons à définir les concepts généraux de plus haut niveau et nous descendons vers les niveaux les plus bas pour définir les concepts spéciaux.

Dans ce contexte, nous avons déterminé cinq concepts généraux : Food, Microbe, Factor, Symptom et Therapy.

Par exemple, à partir du concept Food, nous définissons ses sous-concepts suivants : perishable-food, medium-food-spoilage, slow-food-spoilage, et water. Et à partir du sous-concept slow-food-spoilage, on peut définir d'autres concepts tels que, nuts, cereals, legumes, et bien d'autres.

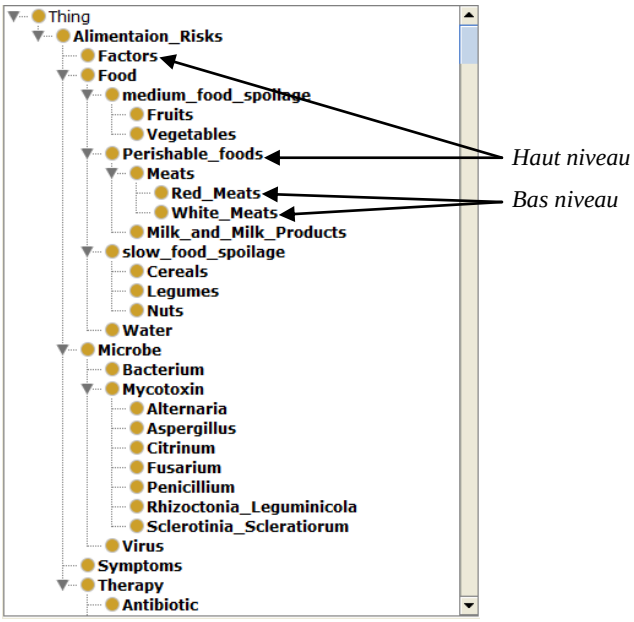


Figure 4.4. Un extrait de la hiérarchie des classes sous Protégé 2000

La construction de la hiérarchie des classes est faite par la définition de différentes relations entre elles. Ces relations peuvent être de différents types :

- *Relation de subsomption IS-A :* ou relation de généralisation/spécialisation qui permet de définir l'héritage entre classes. L'héritage est l'un des mécanismes fondamentaux de l'approche orientée-objet. Il permet de créer des classes à partir de classes existantes. Par exemple, perishable-food est une sous-classe de la superclasse Food, et toutes les instances à définir dans perishable-food seront aussi des instances de la classe Food. La classe perishable-food héritera les attributs et les facettes des attributs de la superclasse Food;
- *Relation Part-of* : ou relation partie-de qui exprime la relation d'appartenance entre deux classes. Elle traduit la relation lexicale Méronymie/Holonomie;
- *Relations entre des classes différentes* : nous définissons nos propres relations liées au domaine de risques alimentaires, par exemple Has-Microbe est une relation sémantique entre la classe source Food et la classe cible Microbe. Ainsi que, la relation Has-Food relie les deux classes précédentes mais dans le sens inverse c'est-à-dire, Microbe est une classe source et Food est une classe cible.

Pour chaque classe, on détermine ses propriétés (slots), ses facettes et ses instances. De plus, nous ajoutons dans chaque superclasse une variable F de type booléen initialisée à non et qui peut prendre la valeur oui lorsqu'on souhaite retourner dans les réponses les images de concept donné.

Par exemple, l'instance ou l'individu Selmonella de la sous-classe Bacteria de la superclasse Microbe est représenté comme suit : familly: enterobacteria, binominal-name: salmonella, discovered-by: Daniel Elmer Salmon, date-of-discovery: 1884, length: 2 to 5 µm, diameters: 0.7 to 1.5 µm, F : non.

4.4.3. Encodage

Après avoir défini la hiérarchie des classes et le modèle conceptuel de l'ontologie ONTARIS, nous devons la représenter par un langage formel en utilisant l'éditeur Protégé 2000, version 4.3. Le choix de cet éditeur est motivé du fait qu'il est extensible et fournit un environnement plug-and-play qui en fait une base flexible pour le prototypage rapide et le développement d'applications [Musen, 15]. Plusieurs plugins pour la visualisation en mode graphique du contenu de l'ontologie, la figure suivante illustre une partie d'ONTARIS par le plugin OWLviz.

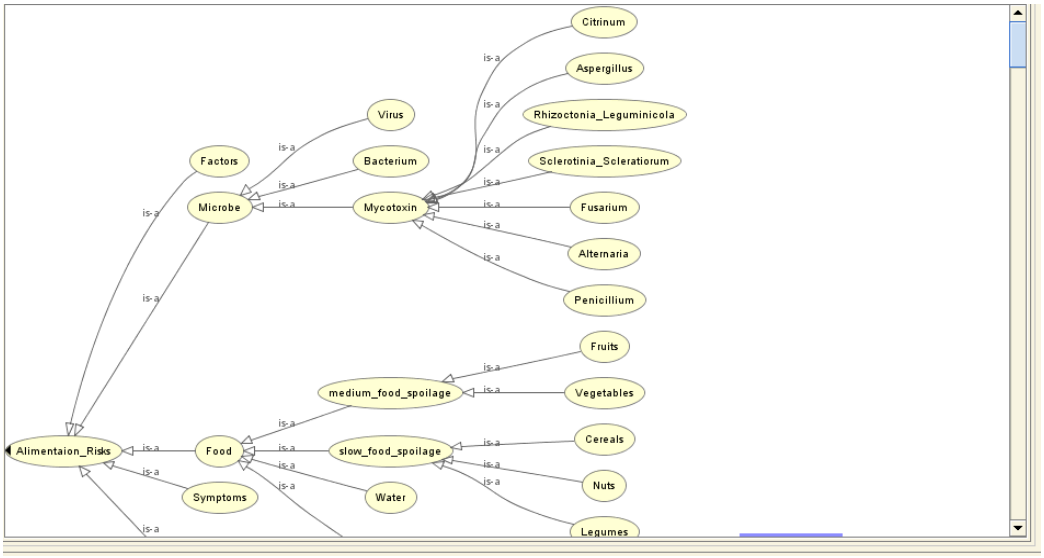


Figure 4.5. Partie de l'ontologie ONTARIS sous Protégé 2000

4.4.4. Evaluation

La dernière étape de construction d'ONTARIS, consiste à évaluer que cette ontologie est correctement construite. Cette évaluation est basée sur deux phases :

- *La vérification terminologique :* qui permet de vérifier la présence de tous les concepts nécessaires pour représenter l'ontologie de domaine de risques alimentaires. Elle doit prendre en compte l'avis des spécialistes de domaine qui n'ont pas participés à la construction de l'ontologie. Dans cette phase, on peut y appliquer des opérations d'ajout et de suppression dans le contenu d'ontologie;
- *La vérification technique :* par l'utilisation des requêtes SPARQL et le raisonneur Jena on peut vérifier la consistance et la bonne structuration des classes dans la hiérarchie.

Finalement, nous devrons utiliser l'ontologie ONTARIS dans notre système de médiation sémantique pour le traitement du problème d'hétérogénéité sémantique dans des bases de données multimédias et hétérogènes.

5. Niveaux d'hétérogénéité sémantique

L'intégration sémantique des sources de données hétérogènes représente une solution prometteuse pour faire face aux problèmes d'hétérogénéité sémantique lors de l'exploration de données multimédias. L'approche de médiateur est l'une des approches d'intégration de

données qui vise à offrir un accès unifié aux données sans besoin de connaître leurs sources d'origine. De ce fait, nous proposons dans ce travail une nouvelle approche de médiation sémantique à base d'ontologies pour l'exploration efficace des bases de données hétérogènes et multimédias. En effet, nous avons situé les problèmes d'hétérogénéité sémantique aux deux niveaux [Aggoune, 17]: hétérogénéité niveau requête et hétérogénéité niveau sources de données.

5.1. Hétérogénéité niveau requête

L'hétérogénéité niveau requête représente les problèmes d'accès et d'exploration des sources de données hétérogènes via des requêtes. Elle est souvent liée à différentes expressions définies par des utilisateurs pour exprimer la même requête et aussi au problème d'ambiguïté du sens des mots composant cette requête. Cette dernière doit être enrichie par des ressources sémantiques et linguistiques pour clarifier sa signification exacte.

En effet, utilisation conjointe d'ontologie comme ressource sémantique et le WordNet comme base linguistique et lexicale nous ont permis de traiter ces différents problèmes d'hétérogénéité au niveau de la requête d'utilisateur. Ainsi, utiliser des composants graphiques pour exprimer la requête d'utilisateur plutôt qu'un langage naturel, permet d'alléger les conflits sémantiques et faciliter la formulation de requête.

De plus, le système de médiation sémantique doit offrir un langage commun et standard pour interroger des bases de données multimédias et hétérogènes. Dans ce contexte, nous utilisons le langage SPARQL pour exprimer des requêtes sémantiques suivant le vocabulaire de l'ontologie partagée ONTARIS. Cette ontologie représente le schéma global du système de médiation et donc un support d'interrogation sémantique qui permet de guider l'utilisateur pour exprimer ses requêtes. En effet, avec le langage SPARQL, nous n'avons pas besoin de connaître a priori la structure et le contenu des données pour pouvoir les interroger. De ce fait, SPARQL permet d'interroger n'importe quel composant d'un triplet qui a la forme Sujet-Prédicat-Objet.

5.2. Hétérogénéité niveau sources de données

Dans ce niveau d'hétérogénéité sémantique, les sources de données représentent la principale cause d'existence des problèmes d'hétérogénéité sémantique quand on a besoin de faire une exploration de données sur ces sources. Diverses représentations de données (images, textes, vidéos et sons), différentes définitions pour décrire les mêmes données, plusieurs modèles pour modéliser les données et bien d'autres, tous ont conduit aux problèmes d'hétérogénéité sémantique au niveau de sources de données.

Dans notre travail, nous avons créé six bases de données multimédias (BDMM) définissant le domaine de risques alimentaire. Ces sources sont hétérogènes ce qui empêche l'accès à leurs données. La section suivante présente notre approche de traitement du problème d'hétérogénéité sémantique pour l'exploration des bases de données multimédias.

6. Approche de médiation sémantique pour l'exploration des BDMM hétérogènes

Les bases de données multimédias (BDMM) sont absolument hétérogènes selon plusieurs critères qu'ils s'articulent autour des problèmes d'hétérogénéité sémantique. Explorer ces sources de données sans aucun problème d'hétérogénéité sémantique est l'un des verrous scientifiques que nous devons étudier et de les traiter. Ainsi, la représentation et l'exploitation du contenu sémantique des BDMM nous ont permis de réduire l'impact nuisible de l'hétérogénéité sémantique.

Il s'agit dans ce cas d'une approche fondée sur les ontologies pour le traitement du problème d'hétérogénéité sémantique lors de l'exploration des BDMM. En effet, nous proposons une nouvelle approche de médiation sémantique par hybridation en utilisant l'ontologie ONTARIS comme ontologie partagée du système de médiation sémantique et les ontologies virtuelles décrivant la sémantique des BDMM.

L'approche de médiation sémantique dédiée à l'exploration des BDMM hétérogènes consiste à considérer l'exécution de la requête d'utilisateur via le médiateur comme un ensemble des matchings ou des appariements entre les éléments composant cette requête et ceux des ontologies virtuelles.

Le but de cette approche est de procéder un processus d'appariement entre la requête exprimée en termes du vocabulaire de l'ontologie partagée ONTARIS et les ontologies virtuelles propres à chaque BDMM plutôt d'effectuer un ensemble de transformations de la requête selon le vocabulaire du langage de source (SQL ou OQL). L'avantage de processus d'appariement est la capacité d'extraire toutes les relations sémantiques qui ne sont pas présentées dans le modèle relationnel ou orienté objet des bases de données.

En effet, aucune réécriture de la requête initiale en termes de vues, ni la traduction des requêtes par les adaptateurs. De ce fait, à l'arrivée de la requête initiale Q exprimée en SPARQL suivant le vocabulaire d'ONTARIS, les adaptateurs doivent appliquer le processus d'appariement entre les éléments d'ONTARIS qui décrivent la requête Q et les éléments des ontologies virtuelles associent à chaque BDMM. Ce processus d'appariement est basé sur l'utilisation de deux mesures de similarité sémantique entre ces éléments : similarité de Wu et Palmer (wup) [Wu et Palmer, 94] et similarité cosinus (cosine similarity) [Salton, 71].

Pour mener à bien le processus d'appariement, l'ontologie lexicale WordNet est mise à la disposition du système de médiation sémantique. Le WordNet a pour objectif de représenter les relations lexicales entre les lexèmes de la langue anglaise. Il permet de désambiguïser le sens des mots en se basant sur la hiérarchie des concepts et l'application d'une mesure de similarité entre concepts.

La figure suivante illustre le déroulement de l'approche proposée. Ce dernier sera défini par huit étapes du 1 jusqu'à 8.

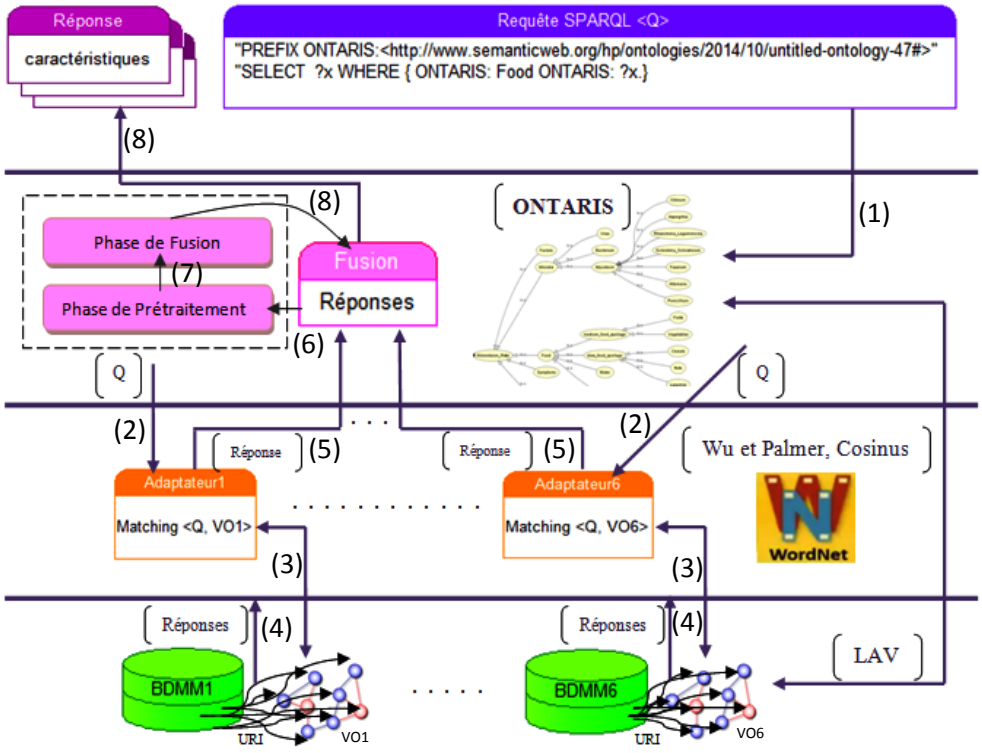


Figure 4.6. Schéma illustratif de l'approche proposée

Le traitement du problème d'hétérogénéité sémantique niveau requête est guidé par l'ontologie partagée ONTARIS qui précise le sens des concepts du domaine de risques alimentaires. La requête de l'utilisateur est alors, une requête sémantique en SPARQL, composée des concepts d'ONTARIS et elle a été construite via la sélection des composants graphiques de l'interface principale (étape (1)). Avec l'utilisation du langage SPARQL comme langage standard et unifié, nous n'avons pas besoin de connaître a priori la structure et le contenu des données pour pouvoir les interroger.

Dans le cas où la requête initiale est une image, il devra nécessaire d'associer à cette image des métadonnées composant des concepts d'ONTARIS permettant de décrire le sens de cette image avec la mise à oui de la valeur de variable F liée à la métadonnée de l'image requête. Dans la couche adaptateurs, chaque adaptateur doit vérifier la valeur de F afin de retourner l'image correspondante à partir du processus d'appariement requête-ontologie virtuelle (étapes (2) et (3)). Le résultat de ce processus est un ensemble de correspondances entre les concepts de la requête et ceux de l'ontologie virtuelle (étape (4)). En effet, chaque adaptateur restitue leurs réponses et il doit les renvoyer à la couche médiateur pour fusionner toutes les réponses des adaptateurs (étape (5)).

La fusion des réponses consiste à attribuer une représentation normalisée selon le vocabulaire de l'ontologie partagée ONTARIS. Une phase de prétraitement (étape (6)) a pour objectif d'éliminer les réponses redondantes qui peuvent être existées dans plusieurs BDMM. L'élimination des réponses vise à supprimer les synonymes des concepts en utilisant le WordNet. Par la suite, une phase de fusion (étape (7)) qui consiste à regrouper les réponses en

utilisant l'union entre elles et les organiser selon les concepts de la requête. Finalement, les réponses normalisées seront retournées à l'utilisateur (étape (8)).

Le traitement du problème d'hétérogénéité sémantique niveau sources de données est accompli par la représentation unifiée du contenu sémantique de sources de données en utilisant les ontologies virtuelles. Ces ontologies permettent de faciliter l'intégration sémantique de données des BDMM. Avant de présenter notre processus d'appariement dédié à l'intégration sémantique de données, nous décrivons tout d'abord les deux mesures de similarité sémantique utilisées dans ce processus.

6.1. Mesures de similarité sémantique utilisées

Nous rappelons que notre approche est basée sur le processus d'appariement entre les concepts de la requête et ceux de l'ontologie virtuelle. Afin d'effectuer l'appariement entre ces concepts, une ou plusieurs mesures de similarité sémantique doivent être appliquées. La similarité sémantique entre deux concepts est le calcul du degré de ressemblance ou dissemblances entre eux [Salton, 71]. De ce fait, les concepts doivent être pondérés par un poids en se basant sur leurs relations de la hiérarchie des concepts dans le WordNet. Dans notre travail, nous avons adopté deux mesures de similarité sémantique entre concepts : la similarité de Wu et Palmer et la similarité cosinus. Utilisation simultanément de ces deux mesures est prometteuse pour assurer l'appariement de qualité.

6.1.1. Similarité de Wu et Palmer

Zhibiao Wu et Martha Palmer ont présenté dans leur article intitulé « *Verb semantics and lexical selection* », une nouvelle mesure de similarité sémantique baptisée *wup* entre concepts dans une ontologie restreinte aux liens taxonomiques [Wu et Palmer, 94]. La mesure *wup* est basée sur le calcul dus profondeur d'un concept (*depth*) qui est représentée par la longueur d'arc entre la racine et ce concept [Wu et Palmer, 94]. Cette mesure est utilisée pour la traduction automatique entre les verbes adaptés aux deux langues anglaise et chinoise [Wu et Palmer, 94]. La mesure *wup* est donnée par la formule suivante :

$$\mathrm{Sim}_{wup}(C1,\ C2) = \frac{2 \times \mathrm{depth}\ (C)}{\mathrm{depth}\ (C1) + \mathrm{depth}\ (C2)}$$

Avec C est le plus proche concept commun entre les concepts C_1 et C_2 (ou le plus petit généralisant *PPG*). Le depth (C) est le nombre d'arc qui séparent C de la racine. Dans le cas où C est la racine, la valeur de depth (C) soit 0. Ainsi, le depth(C_1) (ou depth(C_2)) est le nombre d'arcs qui sépare C_1 (ou C_2) de la racine en croisant le concept C. Les valeurs de wup sont comprises entre 0 et 1. La valeur 0 indique que deux concepts ne sont pas similaires tandis que la valeur 1 indique la similarité.

Prenons l'exemple illustré dans la figure suivante pour calculer la similarité sémantique *wup* entre le concept vegetable et le concept fruits [Aggoune, 17].

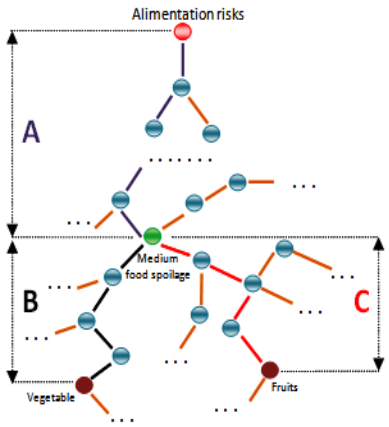


Figure 4.7. Exemple de la hiérarchie de classe de l'ontologie ONTARIS

Alors, la similarité sémantique entre vegetable et fruits est la suivante:

$$\mathrm{Sim}_{wup}(\text{vegetable, fruits}) = \frac{2\times \text{depth}\left(\text{medium}_{\text{food}_{\text{spoilage}}}\right)}{\text{depth (vegatable)} + \text{depth (fruits)}}$$

Avec: depth(medium_food_spoilage)=A, depth(vegetable)=B+A, et depth(fruits)=C+A. Et ainsi la similarité *wup* devient: $\mathrm{Sim}_{wup}(\text{vegetable, fruits}) = \frac{2\times A}{B+C+2*A}$.

La mesure de similarité de Wu et Palmer a l'avantage d'avoir un temps d'exécution court et donne une bonne performance par rapport les autres mesures. Néanmoins, elle retourne des valeurs erronées quand un concept est composé de plus de deux mots comme date-of-discovery comprend deux mots date et discovery.

Pour cette raison, nous avons combiné cette mesure avec d'autre mesure de similarité dont ses valeurs sont entre 0 et 1 et qui est fondée sur la représentation vectorielle de concept. En effet, nous utilisons la mesure la plus populaire, il s'agit la similarité cosinus décrite dans [Salton, 71].

6.1.2. Similarité Cosinus

La similarité Cosinus est l'une des similarités les plus souvent utilisées dans les domaines de la recherche d'informations (RI) et le traitement automatique des langues naturelles (TALN). Elle est notamment utilisée dans les applications de Text Mining pour mesurer la ressemblance entre deux segments de texte [Li, 13].

La similarité cosinus entre deux concepts C_1 et C_2 est une mesure de cosinus de l'angle formé par le vecteur X du C_1 et le vecteur Y du C_2 selon la formule suivante [Salton, 71]:

$$Sim_{cosinus}(X, Y) = \cos(X, Y) = \frac{X.Y}{\|x\|^2 \times \|y\|^2}$$

Les vecteurs X et Y contiennent les poids des mots composant C_1 et C_2 respectivement et ||X|| est la magnitude de X. Le calcul du produit scalaire X.Y est donné par l'équation suivante:

$$X.Y = \sum_{i=1}^{n} x_i.y_i = x_1.y_1 + x_2.y_2 + x_3.y_3 + \cdots + x_n.y_n.$$

La magnitude de X est : $\|X\| = \sqrt{\sum_{i=1}^{n} x_i^2} = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2.$

Alors, deux concepts sont similaires cela implique que l'angle formé de leurs vecteurs soit petit et la valeur de similarité cosinus est grande. En effet, la similarité entre un concept et lui-même est égale à la valeur de cosinus de l'angle 0 c'est-à-dire Cosinus (0)=1.

La similarité cosinus est très utile pour mesurer la ressemblance entre les concepts incomplets ou syntaxiquement erronés comme par exemple, les mots Aspergill et Aspergillus dont la similarité cosinus est égale à 0.934 par contre avec la similarité *wup* est 0.222. Néanmoins, cette mesure donne un mauvais résultat quand les deux concepts sont syntaxiquement proches et sémantiquement loin, par exemple, la similarité cosinus Apple et Applet est égale à 0.935 (ou 93.5% similaire) par contre avec la mesure de Wu et Palmer est égale à 0.190. De ce fait, la similarité cosinus est inutile pour mesurer la similarité entre synonymes. C'est pour cette raison, on combine avec la similarité de Wu et palmer qui donne un meilleur résultat pour calculer la similarité sémantique entre synonymes.

Par conséquent, la mesure de similarité de Wu et Palmer et celle Cosinus sont indispensables pour assurer l'appariement ou la mise en correspondance entre les concepts de la requête du système de médiation sémantique et les concepts des ontologies virtuelles. La section suivante décrit notre processus d'appariement requête et ontologies virtuelles.

6.2. Traitement de requêtes via le processus d'appariement

Exécuter la requête venue de couche médiateur via les adaptateurs revient à effectuer l'appariement entre le contenu de cette requête et l'ontologie virtuelle propre à chaque source de données multimédias (BDMM). Notre processus d'appariement se déroule en quatre phases d'appariement ordonnées selon cet ordre: appariement concepts, appariement instances, appariement propriétés et appariement relations. De ce fait, le résultat d'une phase sera utilisé comme entrée de la phase suivante. Le résultat final de ce processus est un ensemble de correspondances qui répond à la requête d'utilisateur. Ainsi, la prise en compte de l'ontologie lexicale WordNet est indispensable dans notre processus d'appariement.

Le traitement de requête d'utilisateur via l'adaptateur i avec i allant de 1 jusqu'à 6 (le nombre des BDMM existantes dans la couche sources de données) est donné par l'algorithme suivant [Aggoune, 17]:

```
Algorithme Traitement de requête au niveau de l'adaptateur i
Entrées : Q : requête;
          VOi : l'ième ontologie virtuelle associée à la ième BDMM
Début
  Décomposer Q : { N : liste d'instances.
                   R : liste des relations.
                   X : liste des variables.
  AC, AI, AP, AR : listes initialisés à vide;
  Pour chaque élément Ni de N faire
    Début
      C: est un ensemble des classes d'ONTARIS contenant l'instance Ni.
      Pour chaque élément Cj de C faire
        AC:= AC+ appariement-Concept (Cj, VOi);
        AI:= appariement-instance (AC, Ni);
        P est l'ensemble des propriétés du concept dans ONTARIS similaire à AI;
        P' est l'ensemble des propriétés du concept AI dans VOi;
      Pour chaque élément Pk de P faire
        AP := appariement-propriétés (Pk, P');
      Pour chaque élément Ri de R faire
        Début
          AR:= appariement-relation (Ri, AP);
          X= Les instances liés entre Ni et AR;
          Si F de l'élément X est oui Alors
            Début
              Extraire le chemin de l'image à partir de l'attribut Photo de VOi
              Afficher l'image
            Fin
          Sinon sortir;
        Fin pour
      Regrouper les réponses;
      Initialiser F par la valeur Non;
  Fin pour
Fin
Sortie : correspondances entre Q et VOi
```

Algorithme 4.4. Algorithme de traitement de requête via l'adaptateur i

Pour bien illustrer le fonctionnement de cet algorithme, dans la section suivante nous nous utilisons un exemple d'une requête d'utilisateur et nous allons détailler le déroulement de l'algorithme selon les quatre phases d'appariement.

6.2.1. Appariement Concepts

Soit la requête Q écrite en termes d'ONTARIS via le langage SPARQL:

```
SELECT ?x
WHERE {
ONTARIS: Apple ONTARIS: has_microbe ?x.
}
```

Cette requête permet d'afficher tous les microbes existant dans les pommes contaminées. Le déroulement de notre algorithme du traitement de requête par les adaptateurs est comme suit: au niveau de l'adaptateur i, une décomposition de la requête Q est effectuée et elle génère trois sous-ensembles :

- N contient les instances de Q : N= {Apple};
- R contient les relations de Q : R= {has_microbe};

- X contient la variable X des microbes.

À partir du contenu N (c'est Apple), on crée C l'ensemble des classes d'ONTARIS où Apple est l'un de leurs instances. Alors, C = {Food, Medium_food_spoilage, Fruit}. Pour chaque élément de C, on réalise la première phase d'appariement, il s'agit l'appariement concepts.

L'appariement concepts vise à calculer au niveau de l'adaptateur i la similarité sémantique entre les concepts de C et ceux de l'ontologie virtuelle VO_i. De ce fait, pour chaque élément de C par exemple Food, on calcule la similarité sémantique entre Food et toutes les classes de VO_i, puis on sélectionne le concept ayant la plus grande valeur de similarité supérieure ou égale au seuil 0.5. En effet, si plusieurs concepts possèdent la même valeur maximale, on prend tous. La table suivante illustre une partie de résultat d'appariement concepts pour la classe Food et la première ontologie virtuelle (i=1).

Concepts de C	Concepts de VO_1	Similarité	Mesure de similarité utilisée
Food	Microorganism	0.40	Cosinus
Food	Feed	0.92	Wup
Food	Factorize	0.22	Cosinus
Food	Treatment	0.42	Cosinus
Food	Indication	0.38	Cosinus
Food	Fruit	0.46	Wup
Food	Medium_food	0.88	Cosinus
Food	Veggie	0.85	Wup
Food	Vitamin	0.86	Wup
Food	Virus	0.46	Wup

Table 4.3. Un extrait de résultat d'appariement concepts de VO_1

Il est clair que le concept pivot similaire au concept Food est Feed. On procède le même appariement avec les autres classes de C, c'est-à-dire Medium_food_spoilage et Fruit.

Le résultat final de l'appariement concepts est donc AC= {Feed, Medium_food, Fruit}. Cet ensemble sera utilisé comme entré de la deuxième phase d'appariement (appariement instances).

6.2.2. Appariement instances

L'appariement instances consiste à utiliser le résultat de l'appariement concepts AC pour retourner la classe la plus similaire. Cet appariement est basé sur le calcul de similarité sémantique entre les éléments de AC et l'instance N= {Apple} de la requête Q. On aboutit ainsi aux résultats suivants :

Similarité (Apple, Feed)	Similarité (Apple, Medium_food)	Similarité (Apple, Fruit)
0.50	0.52	0.91

Table 4.4. Résultat d'appariement instances de VO_1

À partir de la table ci-dessus, le concept qui correspond l'instance Apple dans VO_1 est Fruit avec la similarité maximale est égale à 0.91≥0.5. Le résultat de cet appariement est inséré dans l'ensemble AI pour l'utiliser dans la phase d'appariement suivante.

6.2.3. Appariement propriétés

Le résultat de l'appariement précédent est la classe Fruit de VO_1 qui est similaire à la classe Fruit de l'ontologie ONTARIS avec la valeur 1 de similarité wup (cette valeur est obtenue dans le premier appariement). A partir de ces deux classes (Fruit de VO_1 et Fruit d'ONTARIS), on mesure l'appariement entre leurs propriétés (slots). Dans ce cadre, P et P' sont deux ensembles des propriétés de Fruit d'ONTARIS et Fruit de VO_1 respectivement. Pour chaque élément P_k de P, on mesure la similarité sémantique entre P_k et tous les éléments de P'. En effet, la propriété similaire à P_k est celle qui dispose la plus grande valeur de similarité. Ensuite, la similarité globale de toutes les propriétés P et P' est la moyenne de toutes les similarités maximales obtenues. Le résultat de l'appariement nommé AP est un ensemble des concepts pertinents de AI. Dans notre exemple, AP=AI= {Fruit}.

Le but de cette phase d'appariement est de vérifier encore une fois la ressemblance entre les concepts (ou classes) de l'ontologie partagée ONTARIS liés à la requête Q et ceux des ontologies virtuelles des sources de données en mesurant la similarité sémantique entre les propriétés de ces concepts.

6.2.4. Appariement relations

La dernière phase d'appariement appelé, appariement relations qui a pour objectif de calculer la similarité sémantique entre les relations sémantiques R de la requête Q et toutes les relations des concepts de l'ensemble AP. Dans notre exemple, R={has_microbe} et AP={Fruit}. La relation sémantique de la classe Fruit de VO_1 la plus similaire à la relation has_microbe est la relation has_microorganism avec sa valeur de similarité cosinus est égale à 0.94. Ensuite, à partir de la relation has_microorganism liée à la classe source Fruit, on détermine la classe cible qui est la classe Microorganism.

Finalement, l'algorithme de traitement de requête retourne toutes les instances de la classe Microorganism, à savoir, Aflatoxin, Altenuene, Rota_virus, Expansin, etc. Dans le cas où la requête utilisateur requiert l'affichage des images des microbes existant dans les pommes c'est-à-dire F de X est égale à oui, notre algorithme doit évaluer la valeur de F et extraire les images des microbes à partir de leurs chemins dans l'attribut photo (le chemin est défini via l'algorithme 4.2).

Après avoir transmettre les résultats au médiateur, il devra nécessaire d'initialiser par la valeur Non du contenu de l'attribut F.

Notre approche de médiation sémantique proposée, doit être validée et évaluée à travers une sérié d'expérimentations. La deuxième partie de ce chapitre présente l'étude expérimentale de cette approche proposée.

7. Expérimentations de l'approche de médiation sémantique proposée

Une première étape d'étude expérimentale est de mettre en œuvre un système de médiation sémantique validant l'approche proposée. Par la suite, l'utilisation des métriques pour l'évaluation des performances du système et comparer notre approche par rapport les travaux similaires.

7.1.Architecture en couches du système SAMER

Nous avons implémenté notre système de médiation sémantique baptisé *SAMER* (**S**em**A**ntic **M**ediation for alim**E**ntation **R**isks). L'architecture du système *SAMER* est divisée en trois couches distinctes : couche médiateur, couche adaptateurs et couche sources de données. La figure suivante montre cette architecture dont le détail sera présenté dans les sous-sections qui vont suivre.

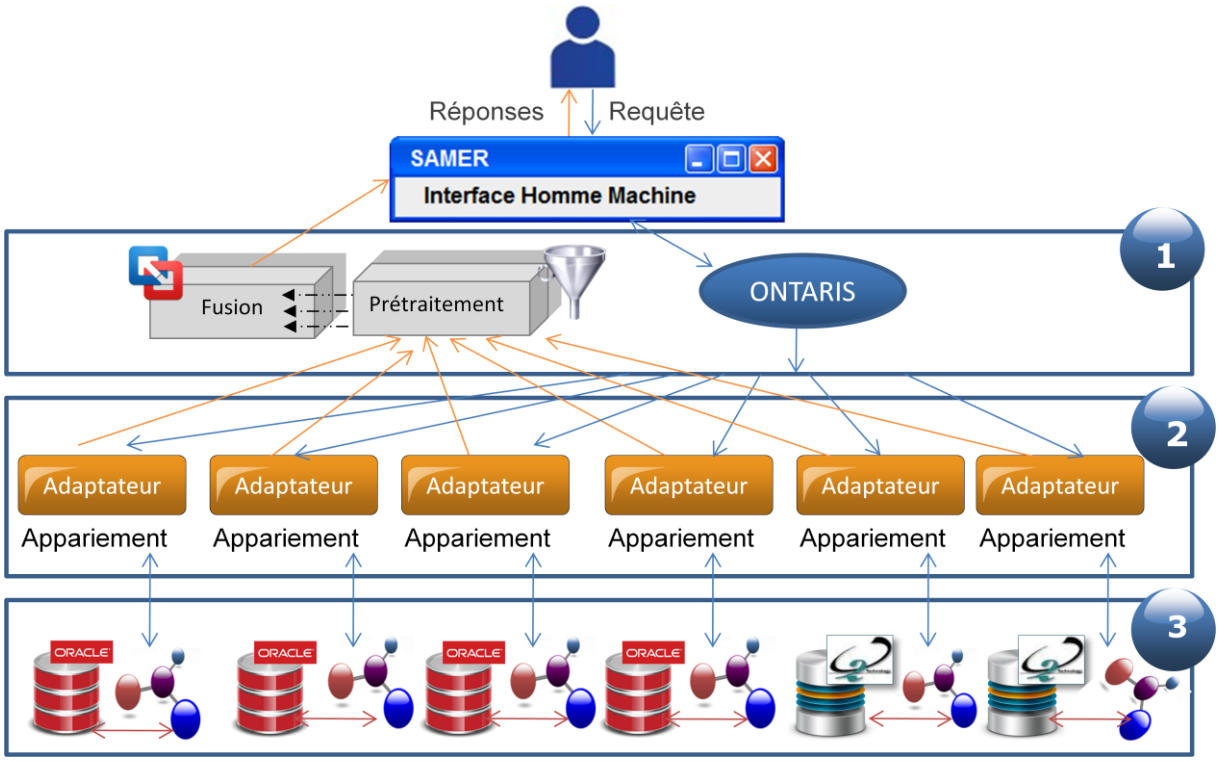


Figure 4.8. Architecture en couches du système SAMER

7.1.1. Couche Médiateur

La couche médiateur permet d'assurer l'accès unifié aux sources de données multimédias. Elle vise à traiter l'hétérogénéité sémantique au niveau de la requête utilisateur par l'utilisation de l'ontologie partagée ONTARIS comme schéma global permettant d'exprimer les concepts composant la requête. Cette dernière sera envoyée à la couche adaptateurs pour y exécuter. En revanche, les résultats obtenus dans chaque adaptateur doivent au préalable être passés par une phase de prétraitement pour analyser et supprimer les réponses redondantes (concepts qui se répètent ou des synonymes (ex. salmonelles et salmonella)). Les résultats de la phase de prétraitement seront par la suite réunis et fusionnés

pour créer des réponses homogènes qui sont retournées à l'utilisateur via l'interface du système *SAMER*.

7.1.2. Couche Adaptateurs

La couche adaptateurs est la couche fondamentale de notre architecture. Elle représente le cerveau du système de médiation sémantique *SAMER*. Chaque adaptateur se charge d'exécuter la requête venue de la couche médiateur suivant notre processus d'appariement requête-ontologie virtuelle. Nous rappelons que ce processus se déroule en quatre phases d'appariement : appariement concepts, appariement instances, appariement propriétés et appariement relations. Le résultat du processus d'appariement est donc une collection de données issues des sources de données hétérogènes. L'ensemble de résultats de tous les adaptateurs doit être envoyé à la couche médiateur pour les fusionner. L'utilisation du processus d'appariement au niveau de chaque adaptateur permet de traiter en parallèle le problème d'hétérogénéité sémantique et cela implique que le temps de réponse est réduit contrairement aux systèmes de médiation existants où la requête doit être réécrite et traduite en termes de langage source.

7.1.3. Couche sources de données

Cette dernière couche comprend les sources de données multimédias et hétérogènes dédiées au domaine de risques alimentaires. Nous utilisons six bases de données multimédias (BDMM) que nous avons présenté auparavant. Chaque BDMM contient environ 800 éléments, pour un total d'environ 7800 éléments et de taille est presque 2.5 gigaoctet de données. Ces données sont représentées par des images au format JPEG (ex. image de bactérie), des textes (description de maladie), des chaînes de caractères (ex. nom de bactérie) et des valeurs numériques (ex. la température, le PH). La table suivante présente un extrait de la table microbe d'une BDMM à base de modèle relationnel.

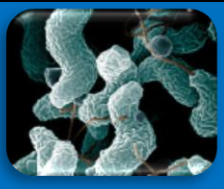
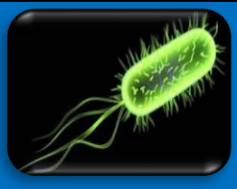
NumM	0001	0002	0003	0004
Binominal-NameM	Salmonella	Pseudomonas	Compylobacter	Escherichia coli
FamillyM	Enterobacteria	Pseudomonadaceae	Campylobacteraceae	Enterobacteriaceae
Date-of-Discovery	1884	1872	1963	1885
Discovered-by	Daniel Elmer Salmon	Schroeter	Sebald et Véron	Theodor Escherich
Length	2 to 5 µm	1.5 to 5 µm	2 to 5.3 µm	0.5 à 3 µm
Category	Bacteria	Bacteria	Bacteria	Bacteria
Diameters	0.7 to 1.5	0.5 to 1 µm	0.6 to 3 µm	0.5 à 5
Photo				
CV	47# Salmonella	47# Pseudomonas	47# Compylobacter	47# Escherichia-coli

Table 4.5. Un extrait de la table Microbe d'une BDMM

Par ailleurs, nous joignons à chaque BDMM sa propre ontologie dite ontologie virtuelle qui a été construite automatiquement à partir de sa BDMM via notre algorithme. Le lien entre la base et son ontologie est faite à travers l'attribut virtuel (ou colonne virtuelle CV) qui

comporte l'URI d'instance dans l'ontologie virtuelle. En effet, notre système SAMER comporte un programme intégré pour la construction d'ontologies à partir d'une base de données plutôt d'utiliser des outils existants comme celui de DataMaster de protégé. Le programme du système SAMER possède un algorithme d'évolution d'ontologie virtuelle qui permet d'assurer la cohérence entre le contenu de BDMM et celui de son ontologie.

7.2. Implémentation du système SAMER

Selon l'architecture du système SAMER, nous devons implémenter tous les algorithmes proposés : écriture de requête, le prétraitement, la fusion, le processus d'appariement, la construction automatique d'ontologie virtuelle et la gestion de type d'attribut, l'évolution d'ontologie virtuelle, et l'affichage des réponses. Pour cela, nous utilisons un micro-ordinateur disposant d'un CPU Intel Core i5 et d'un processeur de 2.50 GHz et ayant 4 Go de RAM avec disque dur de 500 Go. Nous utilisons également le système d'exploitation Windows 7 et l'environnement de développement Eclipse Luna du langage de programmation JAVA.

7.2.1. Outils de développement utilisés

SAMER (SemAntic Mediation for alimEntation Risks) est le système de médiation sémantique que nous avons mis en place en utilisant les outils logiciels suivants :

- ***Eclipse Luna*[20] *:*** est une nouvelle version de l'environnement de développement Eclipse qui a été mis sur le marché en 2014. Il s'appuie principalement sur le langage orienté objet Java[21] de Sun Microsystems. Eclipse Luna inclut le support natif de Java 8 et fournit des plugins, parmi ceux que nous allons utiliser dans SAMER sont : L'API *Jena 2.6.3* pour la création et la manipulation des ontologies, *WS4J 1.0.1* et *JawJaw 1.0.2* pour calculer la similarité sémantique, *Jfreechart 1.0.19* pour l'utilisation des graphiques (histogramme, secteur, barre, etc.).
- ***Protégé 2000:*** est un éditeur des ontologies, développé à l'université de Standford [Noy, 00a] (plus de détail, consulter le chapitre 3). Nous utilisons le langage OWL pour construire nos ontologies. Protégé 2000 dispose des plugins pour la visualisation des ontologies en mode graphique comme, *Jambalaya 2.7.1* et *Graphviz 2.38*.
- ***Wordnet 2.1 :*** est une ontologie linguistique ou encore une base de données lexicales de la langue anglaise (plus de détail, consulter le chapitre 3).
- ***SGBDs :*** comme nous avons dit dans la première partie du chapitre, que nous avons modélisé les BDMM en utilisant deux modèles différents avec deux SGBDs différents : le modèle relationnel par le SGBD *Oracle 10g Express Edition* et le SGBD *O2*, pour la gestion des bases de données orienté-objet.
- ***Outils de traitement d'images :*** nous utilisons *Microsoft Picture Manager*[22] pour redimensionner la taille des images et *PhotoFiltre7*[23] pour effectuer des retouches et des

[20] https://projects.eclipse.org/releases/luna.
[21] https://www.java.com/
[22] https://www.microsoft.com/
[23] http://www.photofiltre-studio.com/pf7.htm

réglages sur l'image (contraste, luminosité, teinte, etc.) avec une large gamme de filtres (effet puzzle, pointillisme, etc.).

7.2.2. Exemples de déroulement de requête d'utilisateur

L'interface principale du système de médiation sémantique SAMER est composée de trois boutons essentiels :

- *Simple query* pour formuler et exécuter une requête simple ne contenant pas des images.
- *Complex query* permet d'exécuter une requête complexe contenant d'une image.
- *ADD Database* permet d'assurer l'évolutivité du système.

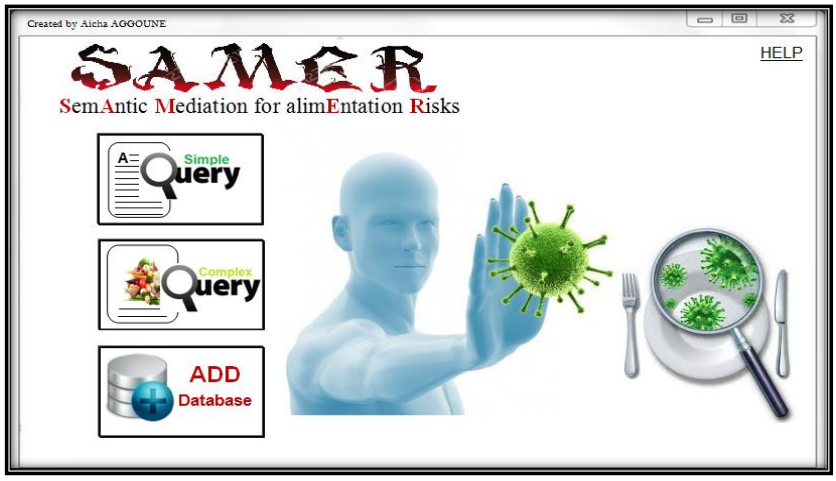


Figure 4.9. Interface principale du système SAMER

L'utilisateur peut rajouter une nouvelle base de données multimédias en spécifiant son modèle (relationnel, orienté-objet), ses entités (classes ou relations), ses attributs et éventuellement ses contraintes et ses méthodes. Ainsi, SAMER doit être capable d'associer un nouvel adaptateur pour cette nouvelle source de données. La figure suivante illustre un exemple d'ajout d'une BDMM à base du modèle relationnel contenant deux tables relationnelles: Bug et Fruit.

Figure 4.10. Interface d'ajout d'une nouvelle BDMM

En ce qui concerne l'exploration des données du SAMER, l'utilisateur a le choix entre l'exploration simple via des requêtes simples ou l'exploration multimédia grâce aux requêtes complexes. Les réponses aux requêtes utilisateurs seront affichées dans une fenêtre séparée.

Nous allons présenter dans la suite de cette sous section, deux exemples de déroulement de requête d'utilisateur : le premier exemple sert à exécuter une requête simple ne contenant pas des objets multimédias (images), par contre le deuxième exemple donne une exécution d'une requête complexe contenant des images.

7.2.2.1.Requête simple

La construction dynamique de requête via l'interface principale du système SAMER est réalisée par une simple sélection des composants graphiques (listes déroulantes, boutons, etc.) sur lesquels va porter la recherche. Utiliser les composants graphiques d'interface homme-machine plutôt d'écrire manuellement la requête, permet d'améliorer notablement l'ergonomie du système et éviter le risque des fautes de frappe et/ou d'orthographe. Nous prenons l'exemple de la requête qui permet d'afficher les aliments contaminés qui peuvent contenir le microbe aflatoxine. La construction de cette requête consiste à sélectionner d'abord la classe Microbe puis choisir l'instance aflatoxine et la relation sémantique Has_food.

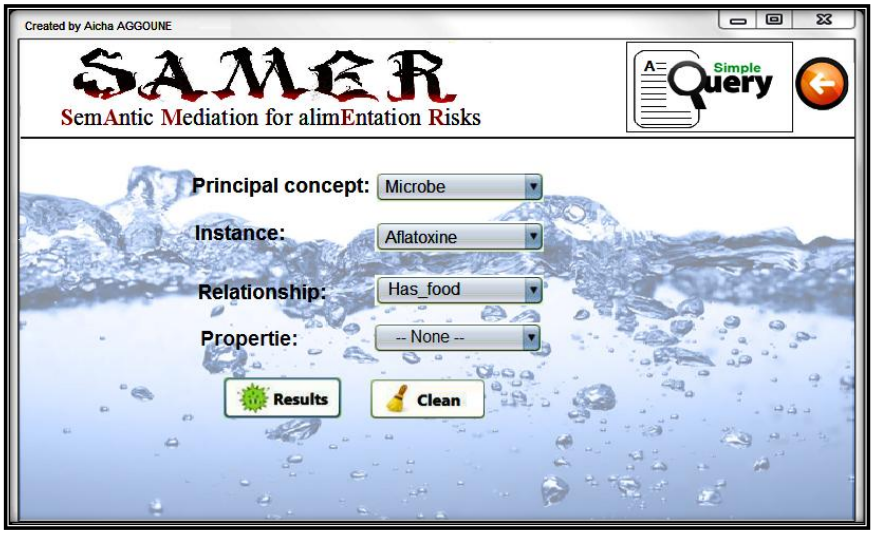


Figure 4.11. Capture d'écran de la formulation d'une requête simple par SAMER

Lorsque l'utilisateur appuie sur le bouton Results, une génération automatique de requête en langage SPARQL puis son exécution via notre processus d'appariement requête-ontologie virtuelle. Afin de montrer correctement l'applicabilité et le fonctionnement du système SAMER, nous avons montré intentionnellement les résultats des quatre phases d'appariement. De ce fait, nous allons prendre d'une part, des captures d'écran de la fenêtre de console d'Eclipse pour visualiser les calculs des similarités sémantiques et d'autre part, des interfaces graphiques illustrant les résultats d'appariement.

En effet, pour appliquer notre processus d'appariement, la première étape consiste à extraire tous les concepts liés à l'instance aflatoxine qui sont : Microbe, Mycotoxin et Aspergillus.

La figure suivante donne un extrait de résultat de l'appariement concepts entre ces trois concepts et les concepts d'ontologie virtuelle.

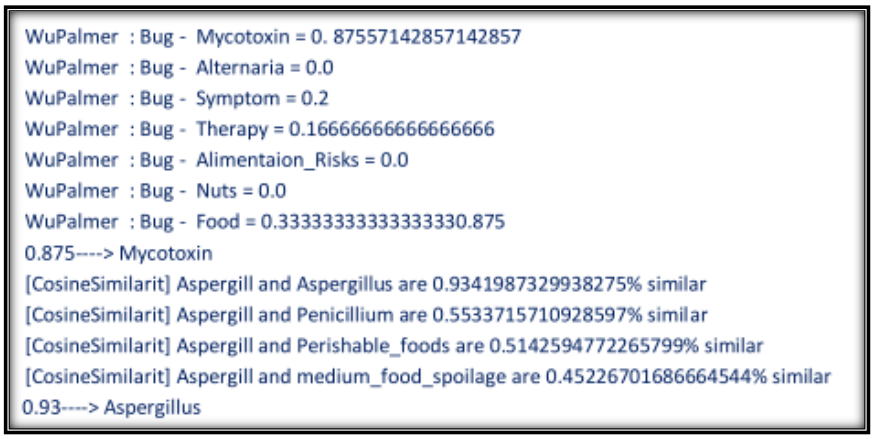


Figure 4.12. Un extrait de résultat de l'appariement concepts

Dans l'appariement instances, le résultat de l'appariement précédent doit être raffiné afin de sélectionner les concepts les plus similaires à la requête utilisateur.

La figure suivante présente l'interface graphique de résultat d'exécution d'appariement instances.

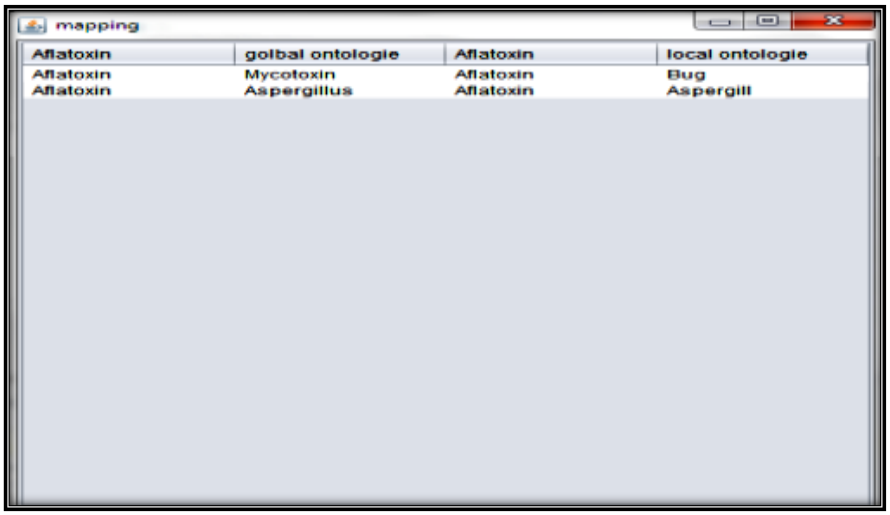


Figure 4.13. Interface graphique de résultat de l'appariement instances

Appariement propriétés permet de déterminer si les propriétés des concepts obtenus de la phase d'appariement précédente sont similaires à celles de l'individu aflatoxine de la requête. Le résultat de ce troisième appariement est illustré dans la figure suivante.

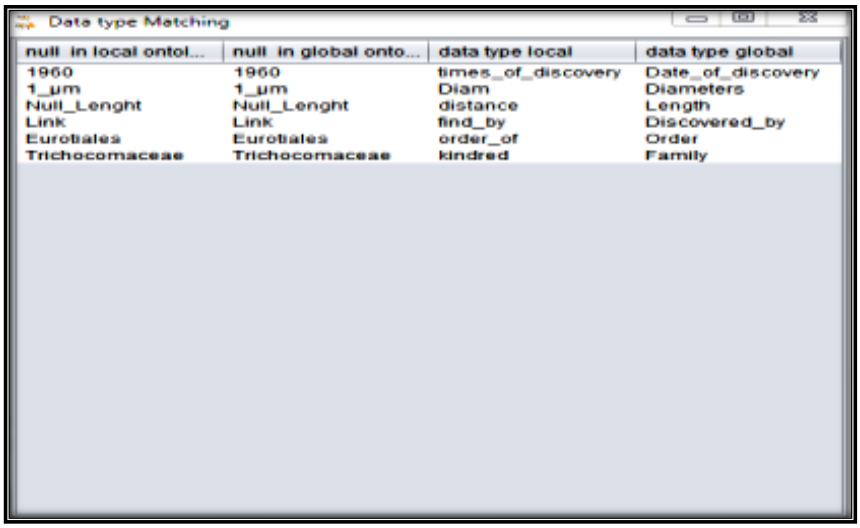


Figure 4.14. Fenêtre graphique de résultat de l'appariement propriétés

La dernière phase est l'appariement relations qui permet de trouver les relations similaires à Has-food et on a obtenu la relation sémantique hasAliment avec une valeur de similarité sémantique est égale à 0.923.

```
WuPalmer   : has_food -   hasAliment = 0.9230769230769231
WuPalmer   : has_food -   hasToken = 0.3076923076923077
WuPalmer   : has_food -   hastraining = 0.2857142857142857
WuPalmer   : has_food -   hasBug = 0.3333333333333333
WuPalmer   : has_food -   hasCause = 0.26666666666666666
```

Figure 4.15. Capture d'écran du résultat de l'appariement relations

Finalement, après le prétraitement et la fusion des réponses de chaque adaptateur, nous visualisons les réponses homogènes comme si c'était une interrogation usuelle d'une seule base de données.

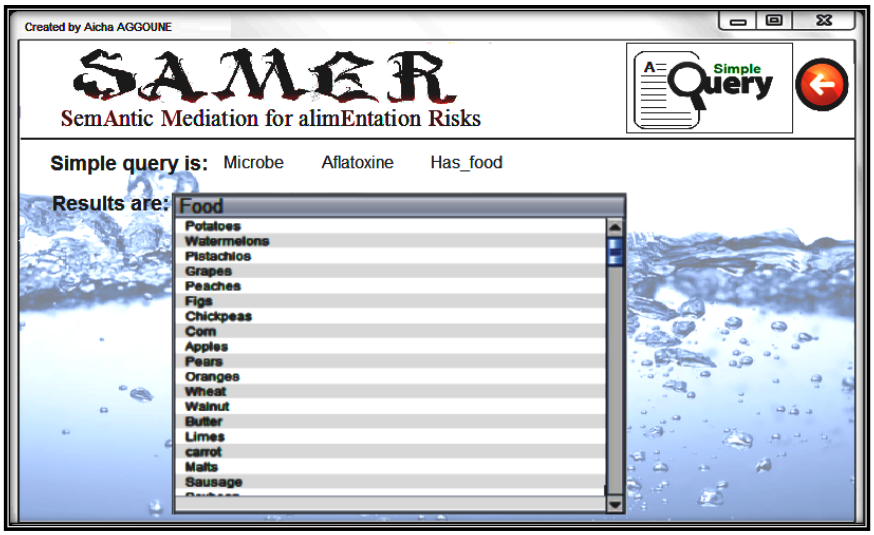


Figure 4.16. Capture d'écran du résultat d'exploration des sources de données hétérogènes

7.2.2.2.Requête complexe

Dans le système de médiation sémantique SAMER, l'utilisateur peut explorer les sources de données hétérogènes à travers une requête complexe composée d'une image afin de récupérer toutes les images qui la ressemblent. L'utilisateur doit fournir quelques concepts de l'ontologie partagée ONTARIS pour décrire la sémantique de l'image requête. En effet, chaque liste déroulante est reliée aux concepts et aux relations sémantiques d'ONTARIS.

La figure ci-dessous montre la formulation d'une requête complexe. L'utilisateur doit donc charger une image qu'il recherche et il propose des concepts d'ONTARIS pour décrire et annoter cette image.

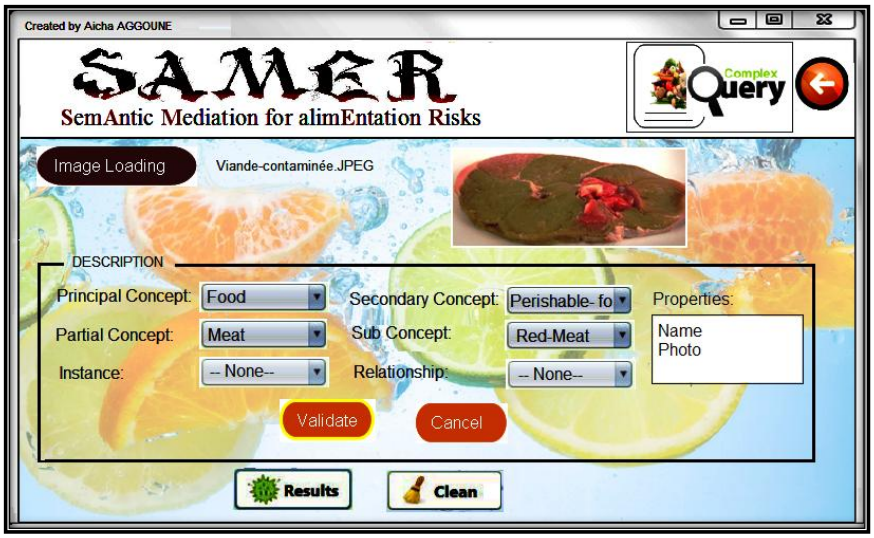


Figure 4.17. Interface de formulation d'une requête complexe

Pour appliquer notre processus d'appariement requête-ontologie, il devra nécessaire de modifier le contenu de l'attribut F du concept principal par la valeur oui (dans notre cas c'est Food). Cette opération est illustrée par la requête SPARQL suivante :

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ONTARIS :<http://www.semanticweb.org/hp/ontologies/2014/10/ontaris#>
    DELETE DATA {
    GRAPH <http://ontaris/Food> {GRAPH ONTARIS: F  "Non"}
    }
    INSERT DATA {
    GRAPH <http://ontaris/Food> {GRAPH ONTARIS: F  Oui}
    }
```

Par la suite nous poursuivons l'algorithme d'exécution de requête dans chaque adaptateur grâce au processus d'appariement et nous présentons ses résultats qui sont illustrés dans la figure suivante.

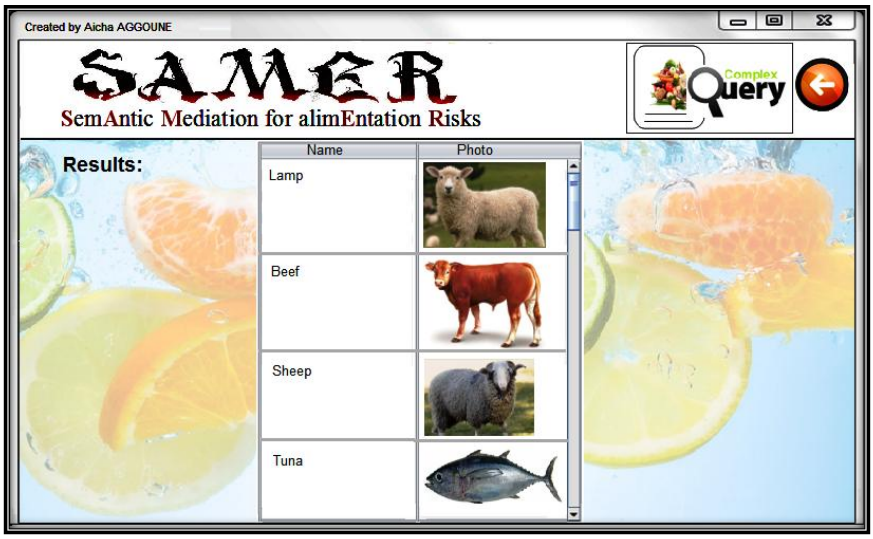


Figure 4.18. Interface de visualisation de résultat d'une requête complexe

7.3.Evaluation des performances

Pour montrer l'intérêt de notre approche de médiation sémantique dans le traitement du problème d'hétérogénéité sémantique des sources de données multimédias et hétérogènes, nous allons évaluer les performances du système SAMER en termes de qualité de réponses retenues et nous discutons les résultats d'évaluation obtenus.

7.3.1. Résultats obtenus

Evaluer la qualité du système d'intégration de données consiste à utiliser les trois mesures d'évaluation du système de recherche d'informations qui sont: la précision, le rappel et la F-mesure. L'adaptation de ces trois métriques aux systèmes d'intégration de données donne les trois formules suivantes :

$$\text{Précision} = \frac{Nombre\ \ de\ donn\acute{e}es\ pertinentes\ \ \ int\acute{e}gr\acute{e}es}{Nombre\ \ de\ donn\acute{e}es\ int\acute{e}gr\acute{e}es}$$

$$\text{Rappel} = \frac{Nombre\ \ de\ donn\acute{e}es\ pertinentes\ \ \ int\acute{e}gr\acute{e}es}{Nombre\ \ de\ donn\acute{e}es\ pertinentes}$$

$$\text{F-mesure} = \frac{2 \times Rappel \times Précision}{(Rappel + Précision)}$$

Par ailleurs, pour utiliser ces trois métriques, nous sommes amenés à créer deux bases de données de test : une base de requêtes contenant 300 requêtes simples et 300 requêtes complexes et une base de réponses qui comporte 2585 données jugées pertinentes.

Ainsi, nous avons mené une expérimentation sur 500 requêtes différentes, simples et complexes. La figure suivante donne les résultats d'évaluation des performances.

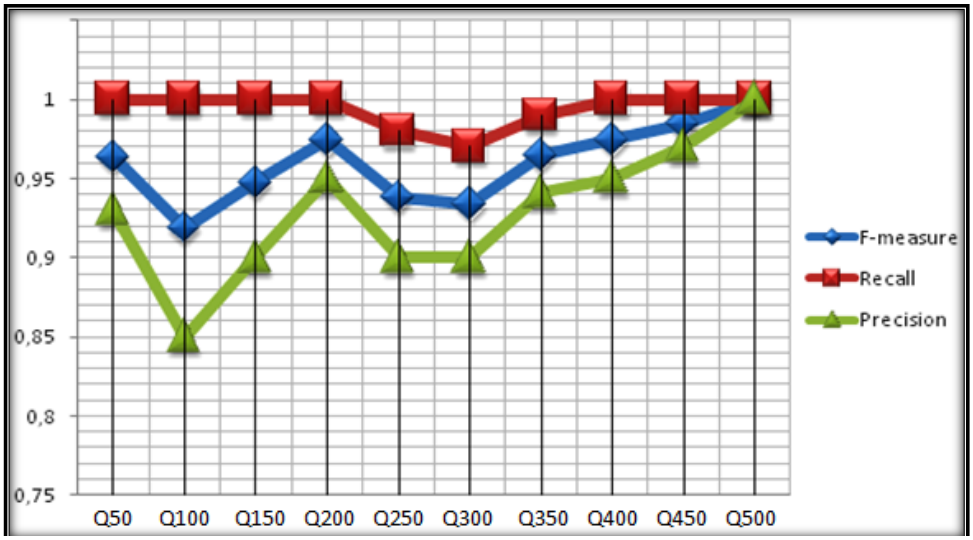


Figure 4.19. Evaluation des performances du système SAMER

7.3.2. Discussion des résultats

Les résultats d'évaluation des performances que nous avons obtenus ont montré que les courbes de rappel, de précision et de F-mesure suivent pratiquement la même allure. Ces résultats indiquent que la performance du système SAMER est très élevée et il est efficace avec une valeur de précision supérieure à 0,95 et une valeur de rappel égale à 1.

Ainsi, la haute précision du système signifie sa capacité d'intégrer seulement les données jugées pertinentes avec un rappel de 100% qui implique que toutes les données jugées pertinentes ont été intégrées. La combinaison de ces deux importantes valeurs de précision et de rappel améliore notablement la valeur de F-mesure. Une analyse des résultats pour plus de 82% des requêtes de la base de test montre l'efficacité et la fiabilité de notre approche de médiation sémantique, notamment le bon fonctionnement du processus d'appariement requête-ontologie qui joue un rôle très important pour le traitement du problème d'hétérogénéité sémantique au niveau des requêtes et des sources de données.

La deuxième expérimentation que nous allons présenter, vise à montrer d'une façon comparative, l'originalité de l'approche proposée d'une part et le bon choix de mesures de similarité sémantique, d'une autre part.

7.4.Comparaisons de l'approche de médiation sémantique

Nous avons mené deux études comparatives de notre approche : une comparaison quantitative et une comparaison qualitative. Le but de la première comparaison est de montrer que les mesures de similarité utilisées ont conduit à des meilleures performances par rapport à d'autres mesures existantes. La seconde comparaison (comparaison qualitative) permet de montrer l'originalité et les performances de notre approche par rapport aux autres approches similaires.

7.4.1. Comparaison quantitative

Afin de valider que les deux mesures de similarité utilisées (Wu et Palmer et similarité cosinus) dans le processus d'appariement requête-ontologie donnent un meilleur résultat par rapport aux autres mesures, nous avons comparé chaque métrique à deux autres. A cet effet, nous nous basons sur les travaux de [Varelas, 05], [Looman, 60] et [Niwattanakul, 13] pour présenter les mesures utilisées pour effectuer cette étude.

De ce fait, la similarité Wu et Palmer a été comparée à deux mesures suivantes: la similarité *Path* et la similarité de Jiang-Conrath [Jiang, 97]. Le choix de ces deux mesures de similarité est motivé par deux raisons : la première est que le principe de calcul est le même que pour la similarité de Wu et Palmer, il est basé sur la profondeur d'un concept dans une ontologie restreinte aux liens taxonomiques, la deuxième est l'intervalle de leurs valeurs est comprise entre 0 et 1. De même, pour la similarité cosinus, nous avons choisi la similarité de Sorensen et la similarité de Jaccard. Ces mesures avaient les mêmes caractéristiques de similarité cosinus (espace vectoriel et l'intervalle des valeurs).

En revanche, pour accomplir cette comparaison quantitative, nous avons utilisé les packages WS4J et JawJaw dans notre système de médiation sémantique. Ainsi, nous avons sélectionné des centaines des paires de concepts pour comparer les valeurs de ces six mesures de similarité précitées.

Dans ce qui suit, nous allons présenter six exemples de paires de concepts ; trois paires sont simples appliquées à la similarité de Wu et Palmer et trois autres sont composées pour comparer la similarité cosinus.

Les résultats de comparaison quantitative ont été implémentés dans une interface graphique en utilisant la bibliothèque Jfreechart 1.0.19 pour afficher les graphiques des résultats de qualité professionnelle.

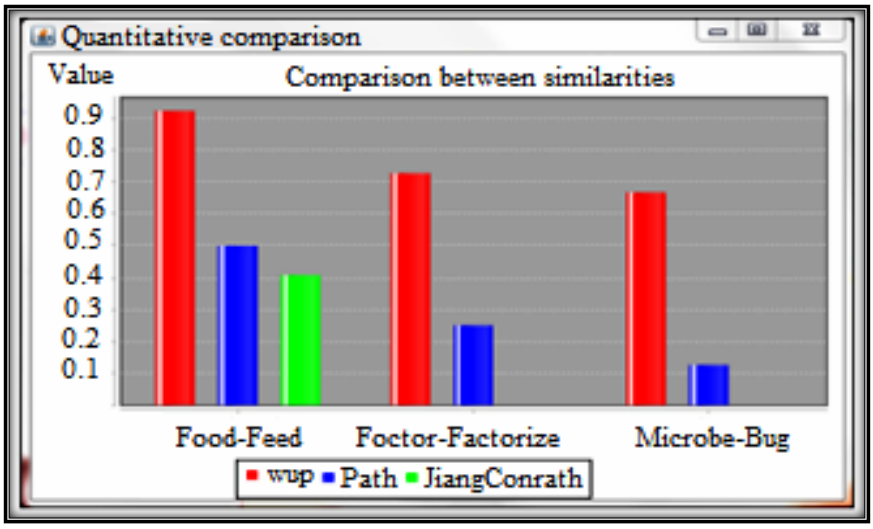


Figure 4.20. Comparaison des résultats entre Wu et Palmer avec Path et Jiang-Conrath

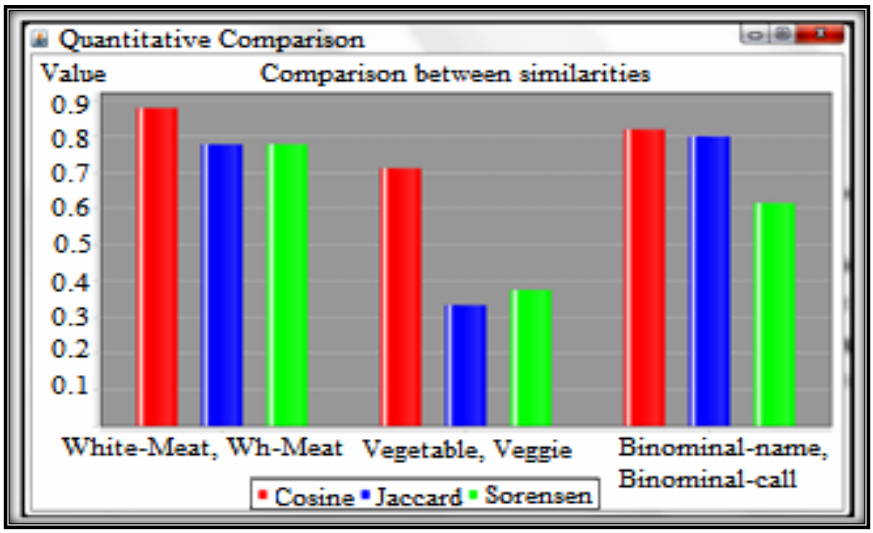


Figure 4.21. Comparaison des résultats entre Cosinus avec Sorensen et Jaccard

D'après les deux figures ci-dessus, nous remarquons que les valeurs de similarité de Wu et Palmer (wup) et celles de similarité cosinus sont très encourageantes par rapport aux autres mesures. Par exemple, la similarité de wup entre deux concepts synonymes food et feed est 0.9231 tandis que dans path et Jiang-Conrath sont respectivement 0.5000 et 0.4078. Les mesures path et Jiang-Conrath donnent des valeurs parfois loin de réalité ce qui dégrade la performance du système, notamment la similarité entre Factor et Factorize qui est nulle avec Jiang-Conrath et 0.2500 avec Path. A partir de ces résultats, nous pouvons dire que l'utilisation de similarité de wup dans notre processus d'appariement permet de traiter correctement les problèmes d'hétérogénéité sémantique aux niveaux de requête et de sources de données. La même chose pour la similarité cosinus, le résultat de comparaison aux mesures Sorensen et Jaccard montre que la similarité cosinus est un bon choix pour comparer deux concepts composés ou syntaxiquement erronés.

En effet, nous pouvons conclure que l'utilisation conjointe de similarité de Wu et Palmer et similarité cosinus permet d'améliorer les performances du système SAMER. Ainsi, la bonne sélection de mesures de similarité utilisées dans le processus d'appariement est l'une des raisons pour laquelle l'approche de médiation sémantique est mieux que les autres approches.

7.4.2. Comparaison qualitative

Afin de montrer l'originalité et la bonne performance de notre travail par rapport les travaux similaires, nous avons mené d'une étude comparative qui va porter sur cinq critères de comparaison : la nature des bases de données à intégrer, le mapping entre schémas, approche de médiation sémantique, Traitement de requêtes et les mesures de similarité utilisées.

Dans cette étude comparative, nous avons également sélectionné trois travaux similaires de domaine de médiation sémantique à base d'ontologie dans le triennal 2011-2013 qui sont :

travail d'Uzdanaviciute et Butleris [Uzdanaviciute, 11], travail de De Giacomo et al [De Giacomo, 12] et le travail de Sultan et al [Sultan, 13].

Nous présentons dans la table suivante le résultat de comparaison qualitative.

Critères / Travaux	Base de données	Approche de mapping	Médiation sémantique	Traitement de requête	Mesure de similarité
Notre travail	Relationnelle et Orientée-objet	LAV	Par hybridation	Quatre phases d'appariement	Wu et Palmer, Cosinus
Uzdanaviciute et Butleris	Relationnelle	LAV	Par hybridation	Algorithme à base des règles métiers	Néant
De Giacomo et al	Relationnelle	GAV	Une seule ontologie	Fonctions logiques	Assertions intentionnelles
Sultan et al	Relationnelle	GAV	Par hybridation	Appariement terminologique	Distance entre chaînes de caractères

Table 4.6. Comparaison qualitative de notre travail avec d'autres travaux

Sur la base de résultat de comparaison qualitative présenté dans la table ci-dessus, nous avons remarqué qu'uniquement notre travail qui permet de traiter le problème d'hétérogénéité sémantique de données multimédias dans à la fois les bases de données relationnelles et orientées-objet, tandis que d'autres travaux ne traitent que les bases de données relationnelles. De plus, durant cette étude nous avons essayé de rechercher dans la littérature des travaux qui traite le problème d'hétérogénéité sémantique dans les bases de données orientée-objet et nous avons trouvé deux travaux : le premier permet d'intégrer des données hétérogènes sans utilisation d'ontologie [Ali, 09], et le deuxième est basé sur l'approche d'entrepôt de données à base d'ontologie [Khouri, 12]. Pour ces deux travaux, on ne peut pas les comparer avec notre travail car ils sont de nature différente, le premier est caractérisé par l'absence d'ontologie et le deuxième est caractérisé par l'intégration via l'approche d'entrepôt de données plutôt que l'approche de médiation.

Par ailleurs, notre travail est basé sur la médiation sémantique par hybridation comme ceux d'Uzdanaviciute et Butleris, et de Sultan et al. Par contre, il est complètement différent au travail de De Giacomo et al qui traite l'hétérogénéité sémantique dans les bases de données relationnelles à travers l'approche de médiation sémantique à base d'une seule ontologie. En revanche, notre travail et celui d'Uzdanaviciute et Butleris appliquent la même approche de mapping LAV.

En ce qui concerne les deux derniers critères de comparaison, nous pouvons dire que la proposition de quatre phases d'appariement entre requête et les ontologies virtuelles, affirme l'originalité de notre travail pour le traitement du problème d'hétérogénéité sémantique dans des sources de données multimédias et hétérogènes. Ainsi, utiliser les métriques Wu et Palmer, et cosinus comme mesures de similarité sémantique et Wordnet comme ressource lexicale permet de distinguer notre travail par rapport les travaux existants.

8. Synthèse des résultats des expérimentations

Pour montrer la fiabilité et l'originalité de notre approche par rapport aux autres approches existantes, nous avons mené deux grandes expérimentations : la première a pour but d'évaluer les performances du système en termes de qualité de données intégrées, et la deuxième sert à faire une étude comparative.

Sur la base des résultats que nous avons obtenus dans la première expérimentation, nous concluons que notre système SAMER a la capacité d'intégrer toutes les données pertinentes et plus rarement de bruit (données intégrées non pertinentes). Ainsi, notre système est très évolutif dans le sens où il pourra être facilement enrichi par une nouvelle base de données et éventuellement un nouvel adaptateur.

Par ailleurs, dans la deuxième expérimentation, nous avons mené deux études comparatives. Une première étude comparative dite quantitative qui permet de comparer les mesures de similarité utilisées dans notre approche avec d'autres mesures existantes. Selon les résultats de comparaison obtenus, la similarité de Wu et Palmer et la similarité cosinus ont été bien choisi en termes de valeur de similarité entre concepts. En outre, utiliser Wordnet comme une ressource lexicale joue un rôle très important pour la désambiguïsation de sens des mots ainsi que pour résoudre l'hétérogénéité sémantique.

Une deuxième étude comparative dite qualitative qui vise à comparer notre travail avec d'autres travaux similaires afin de montrer l'originalité et tirer ses avantages par rapport les autres travaux. Les résultats ont montré l'efficacité de notre approche pour le traitement du problème d'hétérogénéité sémantique dans les bases de données multimédias et hétérogènes. Les données-images sont représentées par un type BLOB dans le modèle relationnel et le type IMAGE dans le modèle orienté-objet. L'interrogation de ces données revient à associer à la requête utilisateur une description sémantique formée d'un ensemble des concepts de l'ontologie partagée ONTARIS. L'approche proposée consiste à appliquer un processus d'appariement entre la requête d'utilisateur et les ontologies virtuelles propre à chaque base de données. Ce processus est défini par quatre phases d'appariement : appariement concepts, appariement instances, appariement propriétés et appariement relations. De plus, notre approche est valide pour deux types de bases de données multimédias BDMM : les BDMM à base de modèle relationnel et les BDMM à base de modèle orienté-objet.

Ces résultats d'expérimentations sont très satisfaisants et ils ont montré l'efficacité de notre approche pour atteindre notre objectif concernant le traitement du problème d'hétérogénéité sémantique pour l'exploration des sources de données multimédias et hétérogènes.

9. Conclusion

Ce chapitre a décrit une première approche proposée pour le traitement d'hétérogénéité sémantique lors de l'exploration des bases de données multimédias et hétérogènes (BDMM). Cette approche de médiation sémantique à base d'ontologies permet d'assurer une intégration sémantique des BDMM en appliquant notre processus d'appariement entre la requête du médiateur et l'ontologie virtuelle propre à chaque BDMM. En effet, nous avons présenté dans

un premier temps trois algorithmes de construction et de gestion des ontologies virtuelles à partir de BDMM et dans un second temps, la construction à partir de zéro de l'ontologie partagée ONTARIS dédié au domaine de risques alimentaires qui représente le schéma global du médiateur.

De plus, nous avons localisé le problème d'hétérogénéité sémantique au niveau de requête et au niveau sources. Dans le premier niveau d'hétérogénéité, la requête doit être guidée par l'ontologie partagée ONTARIS afin d'exprimer sémantiquement les besoins d'utilisateur. Dans le second niveau d'hétérogénéité, une représentation sémantique du contenu des sources via les ontologies virtuelles est essentielle pour assurer l'intégration sémantique de qualité. Cette dernière est basée sur l'application de notre processus d'appariement dans chaque adaptateur. Ce processus est décomposé en quatre phases d'appariement : appariement concepts, appariement instances, appariement propriétés et appariement relations. Il est basé sur l'utilisation conjointe de deux mesures de similarité sémantique : mesure de Wu et Palmer et mesure cosinus.

La deuxième partie du chapitre a été consacré à la présentation des expérimentations permettant de montrer l'utilité et l'efficacité de notre approche proposée. Nous avons présenté l'implémentation du système de médiation sémantique SAMER et nous avons confirmé sa validité, son efficacité et son évolutivité. Les résultats d'évaluation des performances ont montré que notre approche est très performante et elle permet d'intégrer seulement les données pertinentes avec très peu de bruit. De plus, selon les résultats de la comparaison quantitative, nous avons affirmé la bonne sélection des mesures de similarité utilisées dans notre approche. La similarité de wup et la similarité cosinus donnent des valeurs correctes, proche de réalité contrairement aux autres mesures existantes. Ainsi, la comparaison qualitative montre l'originalité de notre approche de médiation sémantique par rapport les travaux similaires. Utiliser les quatre phases d'appariement requête-ontologie permet d'améliorer notablement la qualité du système de médiation sémantique notamment pour le traitement du problème d'hétérogénéité sémantique des bases de données multimédias et hétérogènes.

En résulte que l'utilisation de notre système SAMER est très essentielle notamment dans les industries de production alimentaire pour trouver des solutions aux problèmes liés à la contamination des aliments. Ainsi, il peut être utilisé dans les laboratoires de microbiologie comme support de l'expertise en microbiologie prédictive, tout en tenant compte des conditions de croissance des microorganismes dans les aliments.

CHAPITRE

05

APPROCHE PROPOSÉE POUR L'INDEXATION PERSONNALISÉE DE DOCUMENTS MULTIMÉDIAS: PRÉSENTATION ET EXPÉRIMENTATIONS

Ce dernier chapitre présente notre deuxième approche dédiée au traitement du problème d'hétérogénéité sémantique pour l'exploration de documents multimédias. Ce chapitre est organisé en deux parties principales: la première, fournit une présentation de l'approche d'indexation personnalisée à base du profil utilisateur pour faciliter la recherche d'informations dans un corpus de documents scientifiques et multimédias. La deuxième partie, présente une étude expérimentale pour valider et évaluer l'approche proposée.

1. Introduction

Nous présentons dans ce dernier chapitre, la deuxième contribution qui vise à proposer une nouvelle approche d'indexation de documents multimédias pour le traitement du problème d'hétérogénéité sémantique lors de l'exploration de ce type de données. Il s'agit l'approche d'indexation personnalisée à base du profil utilisateur. Cette deuxième contribution est organisée en deux parties :

La première partie, sert à présenter l'approche proposée par la définition du modèle sémantique d'indexation personnalisée d'une part et le processus d'indexation personnalisée d'autre part. Le modèle proposé est dédié aux documents scientifiques et multimédias qui peuvent contenir plusieurs types de médias (texte, image, audio et la vidéo).

La deuxième partie, décrit l'étude expérimentale de l'outil d'indexation personnalisée *Persodexing* (Personalized indexing) qui a été implémenté en utilisant le langage de programmation Java sous l'environnement de développement Eclipse Luna. Nous allons présenter les interfaces graphiques de l'outil Persodexing et nous évaluerons ses performances en termes du nombre et de qualité d'index générés ainsi que le temps d'indexation personnalisée. Pour cela, nous établirons une comparaison de résultat avec l'indexation classique qui ne prend pas en compte le profil utilisateur dans ce processus d'indexation.

2. Approche d'indexation personnalisée de documents scientifiques et multimédias

Pour combler l'écart entre la nature sémantique des requêtes d'utilisateurs et les documents multimédias, des travaux ont été proposés dans la littérature permettant d'offrir un accès personnalisé à ce type complexe de données par l'utilisation du modèle utilisateur décrivant son profil et l'intégrer dans le processus d'accès aux données afin de mieux répondre à ses besoins en information [Sebe,07] [Zhang, 08]. D'autres travaux sont fondés sur l'indexation sémantique à base d'ontologie permettant de décrire le contenu sémantique de documents multimédias [Bloehdorn, 05] [Mylonas, 08] [Gennaro, 11] [Guo, 14].

Dans cette optique, notre approche proposée consiste à faire appel à la personnalisation de recherche d'informations via le profil utilisateur pour proposer une approche d'indexation personnalisée de documents scientifiques et multimédias d'une part, et à la représentation sémantique de données du domaine de la recherche sémantique via l'ontologie pour définir un modèle sémantique d'indexation personnalisée.

Notre approche proposée assure la prise en compte du profil utilisateur dans le processus d'indexation personnalisée de documents scientifiques et multimédias afin d'améliorer les index de base de documents par l'utilisation des entités complémentaires décrivant à la fois le document et le profil utilisateur. Cette approche présente un modèle sémantique à base d'ontologie pour l'indexation personnalisée de documents multimédias.

L'idée générale de cette approche consiste à utiliser le profil utilisateur comme élément de référence au processus d'indexation personnalisée. Les index de base de documents sont enrichis par des entités complémentaires en exploitant les concepts communs entre la

représentation sémantique du document multimédia et celle du profil utilisateur [Aggoune, 14a].

Cette représentation sémantique est basée sur l'utilisation d'une ontologie de domaine pour décrire à la fois la sémantique des centres d'intérêt d'un utilisateur et la sémantique de documents multimédias. De ce fait, nous avons procédé la même stratégie de construction d'ontologie ONTARIS pour construire l'ontologie **IROnto** (**I**nformation **R**etrieval **Onto**logy) dédiée à la description de domaine de la recherche d'informations.

La figure suivante illustre les principaux concepts d'IROnto en utilisant l'éditeur Protégé2000.

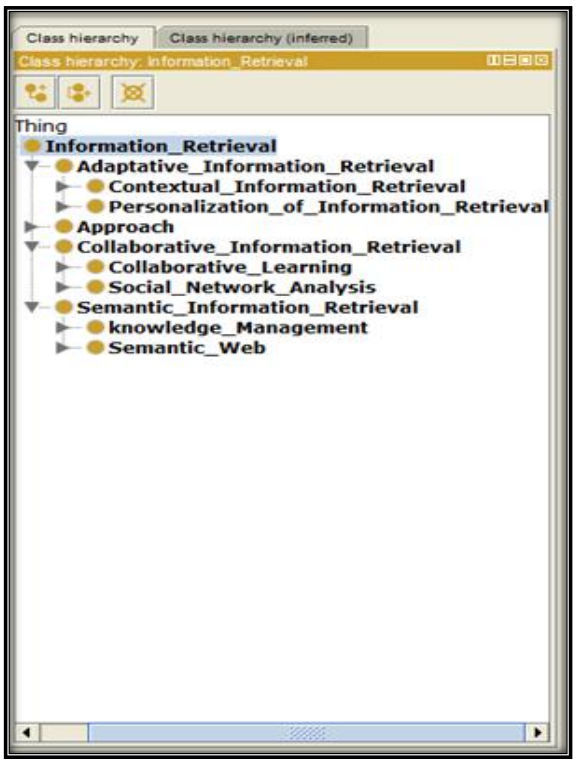


Figure 5.1. Les principaux concepts de l'ontologie IROnto

Dans les deux sections suivantes, nous allons présenter la modélisation sémantique des documents multimédias et celle du profil utilisateur.

3. Modélisation sémantique des documents scientifiques et multimédias

La modélisation sémantique des documents multimédias consiste à exploiter une ontologie de domaine pour représenter le contenu sémantique de ces documents [Guo, 14]. Dans ce travail, nous utilisons notre ontologie IROnto du domaine de recherche d'informations pour décrire la dimension sémantique des documents scientifiques et multimédias.

Nous présentons dans la figure suivante, le diagramme de classes via l'outil Visual Paradigm for UML[24] pour modéliser les documents qui rapportent les publications scientifiques (les articles scientifiques) de domaine de la recherche d'informations [Aggoune, 15a].

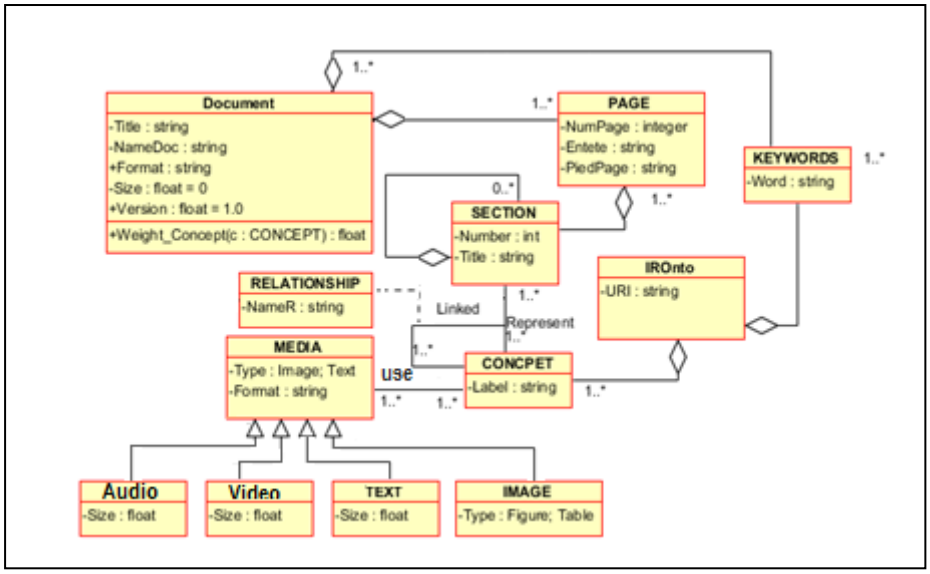


Figure 5.2. Diagramme de classes de modélisation sémantique des documents scientifiques et multimédias

Un document scientifique est identifié par un titre, un nom et caractérisé par un format de stockage avec sa taille et sa version. Il comprend un ensemble de pages et de mots-clés qui reflètent l'objectif du document. Chaque page est composée par au moins une section qui est elle-même peut être composée des sous-sections.

Une section est représentée par un numéro, un titre et un ensemble des concepts liés avec l'ontologie IROnto via l'URI. Ces concepts sont reliés entre eux par des relations lexicales et sémantiques. Ils peuvent être représentés par des objets statiques décrits par des éléments textuels et visuels (images d'une figure ou d'un schéma, des tables) et des objets dynamiques comme des présentations vidéos et enregistrement sonores.

Afin de faciliter l'indexation sémantique de documents multimédias, nous avons ajouté la fonction Weight_concept(C: CONCEPT) qui calcule les poids des concepts liés au document pour sélectionner leur concepts pertinents.

Le poids du concept est obtenu en multipliant les deux mesures statistiques suivantes [Salton, 88]: TF (term frequency) et IDF (Inverse of Document Frequency). Sachant que TF mesure la fréquente du terme t dans le document et IDF mesure la pertinence du terme dans tous les documents du corpus. L'IDF est calculé par log (N/d_f), où N est le nombre total de documents du corpus et d_f est le nombre de documents contenant le terme t.

[24] http://www.visual-paradigm.com

4. Modélisation sémantique du profil utilisateur

La recherche d'informations personnalisée est un type de recherche d'informations qui prend en compte un modèle de l'utilisateur appelé profil pour décrire mieux les besoins en information de l'utilisateur [Dasan, 98].

Le profil utilisateur est une base de connaissance regroupant les caractéristiques utiles pour le comportement du système et les centres d'intérêts personnalisables à chaque utilisateur ou groupe d'utilisateurs [Wahlster, 86].

Nous nous basons sur les travaux de [Ferreira, 01] et [Bouzeghoub, 05] pour définir le contenu du profil utilisateur. De ce fait, le profil utilisateur est généralement composé de données que nous pouvons les regrouper dans deux principaux groupes [Aggoune, 14b]:

- *Les données orientées utilisateur* qui sont les données produites de manière explicite par l'utilisateur à l'aide des formulaires;
- *Les données orientées recherche* qui permettent de déterminer les centres d'intérêt d'utilisateur et ses comportements vis-à-vis le système de recherche d'informations.

Dans le premier groupe de données, l'utilisateur doit faire des efforts pour identifier ses données personnelles telles que l'identifiant (nom utilisateur et mot de passe), le nom, les prénoms, le genre, l'âge, la nationalité, l'email et l'affiliation. Il inclut aussi les activités liées au travail, domaine du travail, du poste occupé et du grade. Ainsi que le niveau d'études avec liste des diplômes obtenus, la spécialité d'étude, le titre du diplôme, mention avec l'année d'obtention.

La collecte de données de ce groupe nécessite un temps pour fournir les données demandées ce qui implique un désintéressement des utilisateurs. Toutefois, ces données sont très utiles au démarrage à froid ; lors de la première connexion d'utilisateur au système.

À l'inverse de données orientées utilisateur, le second groupe de données comprend les centres d'intérêt décrivant les requêtes d'utilisateur, le contexte de recherche avec les niveaux de sécurité de données. Ainsi que les paramètres techniques tels que les systèmes d'exploitation connus, les éditeurs de documents maitrisés et les formats de documents multimédias à rechercher.

Les données orientées recherche peuvent être définies aussi par l'historique de recherche qui est présenté par l'ensemble de documents visités, les liens et les requêtes exécutées.

Toutes ces données peuvent être représentées par différentes représentations du profil utilisateurs [Bouzeghoub, 05] [Gauch, 07] à savoir, la représentation vectorielle des termes pondérés, la représentation sémantique des concepts pondérés d'une ontologie et la représentation multidimensionnelles.

Dans le cadre de notre travail, nous utilisons la représentation sémantique à base d'ontologie pour décrire le contenu du profil utilisateur. Le choix de cette représentation est motivé par le fait qu'elle représente une meilleure solution pour la modélisation du profil utilisateur en prenant en compte les liens sémantiques entre les éléments composants le profil.

La figure suivante présente le diagramme de classes pour la modélisation sémantique du profil utilisateur.

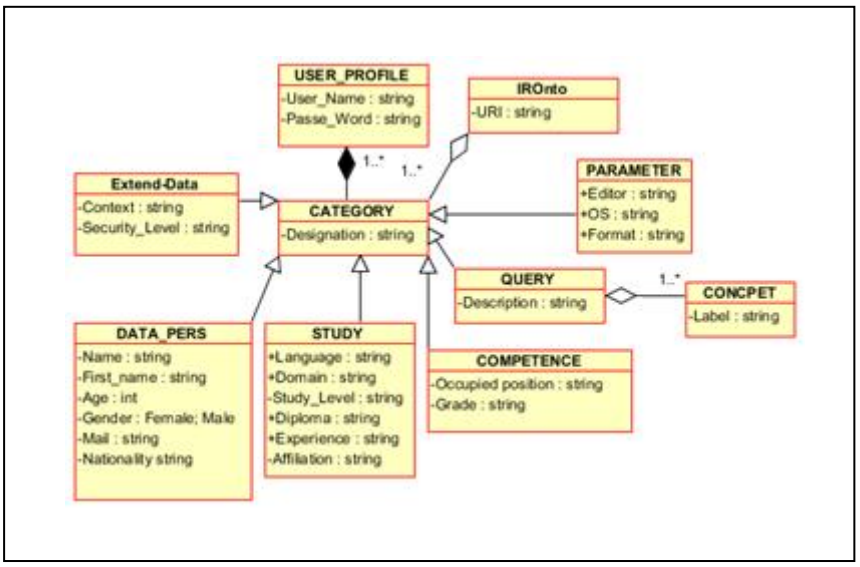


Figure 5.3. Diagramme de classes de modélisation sémantique du profil utilisateur

L'intégration du profil utilisateur dans le processus de recherche d'informations est indispensable pour exprimer correctement ses préférences et ses besoins en information.

Dans la section suivante nous montrons comment le profil utilisateur peut améliorer l'indexation de documents multimédias.

5. Processus d'indexation personnalisée de documents scientifiques et multimédias

Dans le cadre du traitement du problème d'hétérogénéité sémantique dans un corpus de documents multimédias, nous avons proposé une nouvelle approche d'utilisation du profil utilisateur dans le système de recherche d'informations. Cette approche vise à utiliser le profil utilisateur comme élément de référence dans le processus d'indexation appelé processus d'indexation personnalisée plutôt que dans le processus d'accès aux données qui est souvent basé sur la requête de l'utilisateur.

L'indexation personnalisée « *repose sur l'utilisation des index de base (les concepts importants dans le document) et des entités complémentaires en exploitant les concepts communs entre la représentation sémantique du document multimédia et celle du profil utilisateur* » [Aggoune, 14a].

De ce fait, la figure suivante présente le modèle d'indexation personnalisée de documents scientifiques et multimédias qui est obtenu par la mise en relation entre le modèle sémantique des documents multimédias et celui du profil utilisateur.

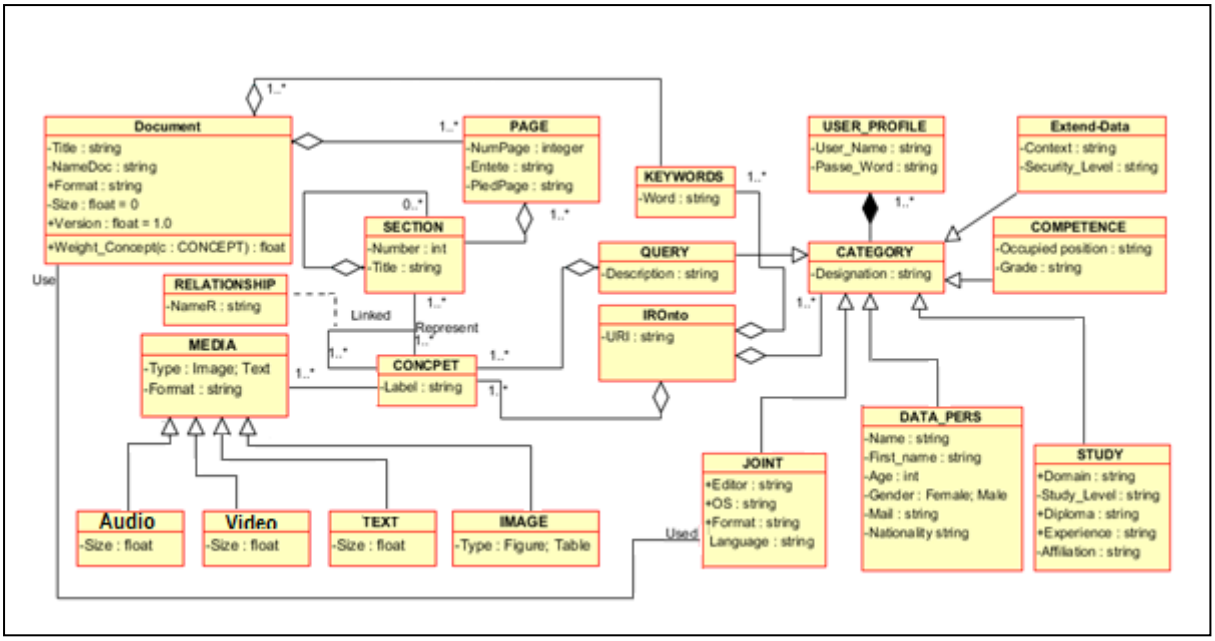


Figure 5.4. Modèle d'indexation personnalisée à base d'ontologie

L'indexation personnalisée dans le modèle présenté au-dessus ne se contente pas de produire des index généraux du document scientifique et multimédia, mais elle génère des index personnalisés propres à chaque profil utilisateur. L'intérêt d'utiliser des index personnalisés plutôt que des index de base, est qu'il peut trouver des concepts non pertinents pour décrire le contenu du document et que ces concepts sont importants pour un utilisateur donné, à titre d'exemples, le format de document (ex. PDF, DOC,..) et les langues maitrisées. Il est donc nécessaire de rajouter ces concepts communs entre l'utilisateur via son profil et le document multimédia dans le processus d'indexation.

Trois éléments communs permettant de lier le profil utilisateur avec les documents multimédias [Aggoune, 14a]:

- L'ontologie IROnto qui permet de décrire à la fois la sémantique de documents scientifiques et multimédias de domaine de la recherche d'informations et la sémantique de données du profil utilisateur;
- La classe CONCEPT est présentée dans les deux modèles sémantiques ; elle permet de définir les concepts contenant un document et la requête d'utilisateur;
- La classe JOINT regroupe toutes les caractéristiques collectives entre le profil et le document (la langue, le format de document, les éditeurs maitrisés, etc.).

À partir du modèle d'indexation personnalisée de documents scientifiques et multimédias, on peut construire les index personnalisés selon le processus d'indexation personnalisée suivant [Aggoune, 15a]:

1. Décomposer le document multimédia en un ensemble des termes;
2. Garder les mots significatifs qui n'apparaissent pas dans l'anti dictionnaire;
3. Les éléments images, tables, audio et la vidéo sont décrites par leurs termes contenant dans leurs titres;
4. Lemmatisation des termes pour donner une certaine normalisation des index;
5. Pondération des index de base suivant la mesure TF×IDF;
6. L'utilisateur a des concepts et des propriétés importants qu'il souhaite les utiliser pour effectuer ses recherches. Ces concepts sont apparus inutiles dans le document, ils doivent donc être lemmatisés, pondérés par un poids est égal à 1 et rajouter à l'ensemble des index.

Le résultat du processus d'indexation personnalisée est donc un ensemble des index personnalisés via le profil utilisateur.

Dans l'étape 03 du processus d'indexation personnalisée, nous avons supposé que dans un document scientifique, tous les objets multimédias (figure, table, vidéo, audio) sont liés par son propre titre décrivant leur sémantique.

Il s'avère nécessaire de valider ce processus dans notre outil d'indexation personnalisée *Persodexing* (*Perso*nalized in*dexing*).

6. Expérimentation de l'outil d'indexation personnalisée *Persodexing*

Rappelons que nous avons proposé une nouvelle approche à base du profil utilisateur pour l'indexation personnalisée de documents scientifiques et multimédias.

Le nouveau type d'indexation proposé est basé sur l'utilisation des termes du profil utilisateur pour enrichir les index de base de document.

L'indexation personnalisée consiste à donner une représentation courte, simplifiée et personnalisée des documents multimédias en exploitant les termes communs entre le document et le profil utilisateur.

Le modèle sémantique d'indexation personnalisée permet en outre d'extraire automatiquement d'autres paramètres utiles d'indexation à partir du profil utilisateur.

6.1. Description de l'outil Persodexing

Nous avons mis en œuvre l'outil Persodexing (Personalized indexing) pour valider l'approche d'indexation personnalisée des documents scientifiques et multimédias.

L'outil développé a été implémenté en Java via l'IDE Eclipse Luna, il utilise la bibliothèque Jena pour parcourir l'ontologie de domaine IROnto (Information Retrieval Ontology) dédiée à la description de domaine de la recherche d'informations.

L'utilisateur peut donc créer son profil via un formulaire comme le montre la figure suivante.

Figure 5.5. Formulaire de création du profil utilisateur

L'utilisateur doit remplir tous les champs du formulaire tels que nom d'utilisateur, mot de passe, email, expériences. Il doit également choisir une propriété dans les listes déroulantes comme les listes de systèmes d'exploitation, de formats du document.

Une fois que l'utilisateur est inscrit à l'outil pour la première fois ou il est connecté s'il a déjà un compte (partie A de la figure 5.6), nous lui affichons l'interface d'indexation personnalisée suivante.

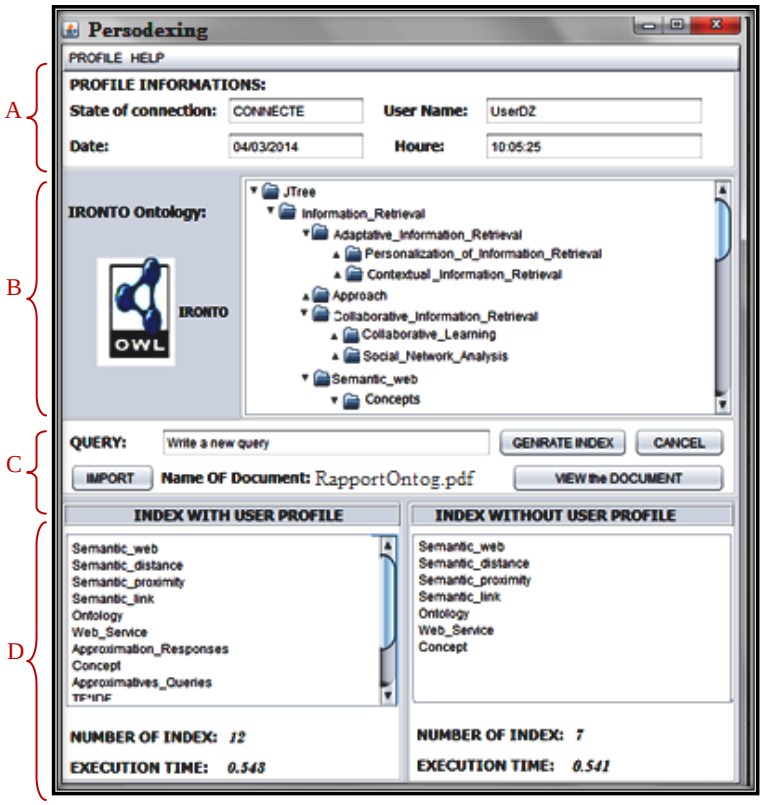


Figure 5.6. Interface principale de Persodexing

La figure ci-dessus présente un exemple d'indexation personnalisée de document multimédia nommé RapportOntog.pdf (cf. partie C de la figure 5.6). L'utilisateur peut parcourir un document existant dans son ordinateur puis il s'appuie sur le bouton GENERATE INDEX pour indexer ce document.

L'outil Persodexing permet aussi d'indexer une requête entrée par l'utilisateur. La partie D de cette figure affiche les résultats de deux types d'indexation : indexation personnalisée et indexation classique. Elle présente ainsi, le nombre des index et le temps écoulé pour générer ces index (plus de détail dans la section suivante).

De plus, l'utilisateur peut visualiser directement le document importé à travers le bouton View Document. Il peut aussi naviguer dans l'arborescence des concepts de l'ontologie IROnto dans la partie B afin de faciliter la formulation de requête à indexer.

6.2. Evaluation des performances de l'outil Persodexing

Nous présentons dans cette section l'évaluation des performances de Persodexing selon le nombre et la qualité des index retournés à l'utilisateur ainsi que le temps d'indexation. Dans ce cadre, nous sommes amenés à effectuer deux modes d'indexation : indexation personnalisée et indexation classique (indexation sans utilisation du profil utilisateur).

De ce fait, cette expérimentation a été faite par 150 utilisateurs différents et 250 documents scientifiques et multimédias. La figure suivante montre les résultats obtenus.

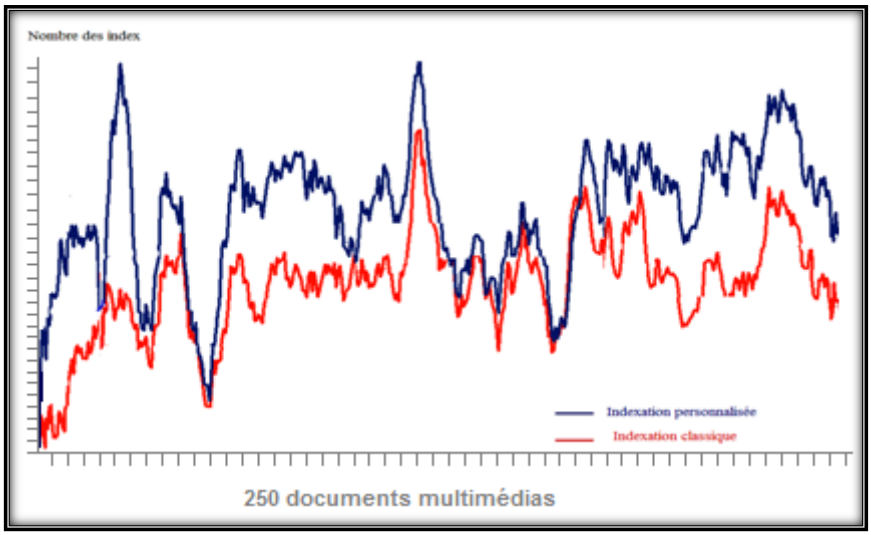


Figure 5.7. Evaluation du nombre d'index selon deux modes d'indexation

D'après la figure ci-dessus, nous pouvons constater que l'indexation personnalisée produit un nombre satisfaisant des index, alors que l'indexation classique produit relativement peu des index qu'ils sont en réalité des index de base de document. Les index résultants de l'indexation personnalisée sont donc des index personnalisés pour un utilisateur donné.

Les index personnalisés vont permettre d'améliorer notablement la qualité d'indexation. De plus, enrichir les index de base par des termes communs entre le document et le profil utilisateur implique systématiquement l'augmentation du nombre des index pour un document donné. Cette augmentation importante permet au mieux de représenter le contenu du document et de faciliter la recherche de documents satisfaisants les besoins d'utilisateur.

Par ailleurs, il devra nécessaire d'évaluer le temps d'indexation pour chaque document indexé. Le résultat d'évaluation est donné dans la figure suivante.

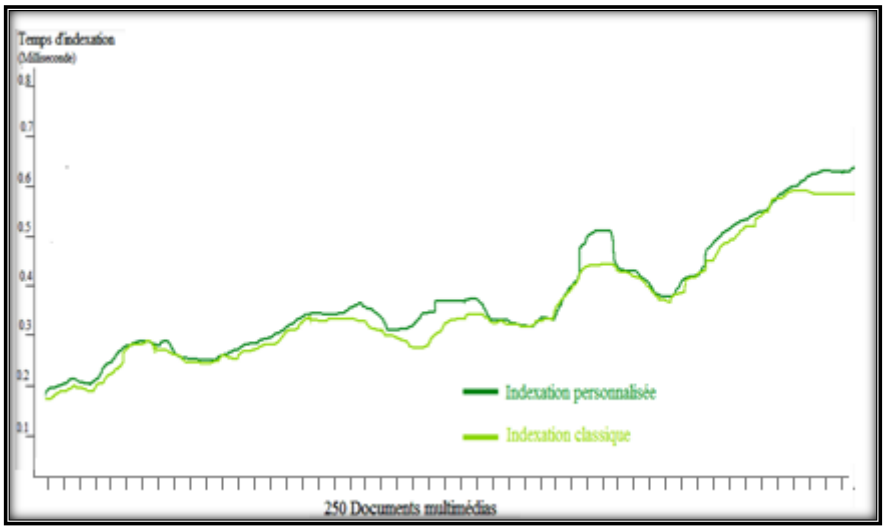


Figure 5.8. Evaluation du temps de réponse selon deux modes d'indexation

Les résultats obtenus montrent clairement que le temps de réponse du processus d'indexation personnalisée est presque le même que dans le processus d'indexation classique.

Par conséquent, notre contribution n'influe pas sur le temps d'indexation et donc au temps de recherche de documents multimédias. En effet, notre deuxième contribution est très efficace et plus adaptée aux documents scientifiques et multimédias.

7. Conclusion

Ce chapitre a été consacré à la présentation de deuxième contribution décrite au cours de cette thèse. Cette contribution donne une nouvelle approche à base du profil utilisateur pour l'indexation de documents scientifique et multimédias. Nous avons proposé un modèle d'indexation personnalisée obtenu par la fusion de deux modèles sémantiques ; le premier dédié à la représentation des documents scientifiques et multimédias et le deuxième permet de représenter le profil utilisateur. Ce modèle vise à enrichir les index de base du document multimédia par des entités complémentaires intéressantes pour un document et un utilisateur donné.

Le résultat du processus d'indexation personnalisée est donc un ensemble des index personnalisés permettant de décrire au mieux les documents multimédias selon les données d'utilisateurs exprimées dans son profil. Ce processus est donc avantageux pour l'utilisateur qui définit correctement son profil.

Dans le reste de ce chapitre, nous avons mené d'une étude expérimentale en comparant notre processus d'indexation personnalisée à base du profil utilisateur avec une indexation classique. Nos résultats sont très encourageants en termes de qualité des index retournés et le temps écoulé pour indexer un document scientifique et multimédia.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Les travaux présentés dans cette thèse se situent dans le contexte du traitement du problème d'hétérogénéité sémantique pour l'exploration des données multimédias qui sont souvent stockées dans des nombreuses sources hétérogènes, conçues indépendamment les unes des autres. L'accès à ces données entraîne des difficultés à exprimer les besoins d'utilisateurs qui peuvent rendre difficile de retrouver les réponses pertinentes. Il est nécessaire d'offrir un accès unifié permettant d'une part, d'explorer aisément ces sources de données et d'autre part de traiter les problèmes d'hétérogénéité sémantique au niveau des requêtes et des sources de données multimédias.

Dans ce cadre, nous avons proposé deux approches distinctes : la première est orientée base de données multimédias, en assurant une médiation sémantique à base d'ontologies, et la deuxième est orientée document multimédia qui présente une indexation personnalisée à base du profil utilisateur. Dans ce qui suit, nous présentons dans un premier temps le sommaire de ces deux contributions et dans un second temps les principales perspectives et orientations pour les travaux futurs.

1. Sommaire des contributions

Nous rappelons que ce travail doctoral s'inscrit dans le cadre de faciliter l'exploration des sources de données multimédias tout en traitant le problème d'hétérogénéité sémantique au niveau de la requête d'utilisateur et les sources de données multimédias. Nous avons présenté un état de l'art qui s'articule autour de trois principaux aspects : aspect représentation et exploration de données multimédias, aspect hétérogénéité sémantique et intégration de données, et aspect sémantique à base ontologies.

Dans le premier aspect, nous avons distingué deux modes de représentation de données multimédias : les documents multimédias et les bases de données multimédias. Nous avons limité notre travail aux données de type image et de texte en utilisant deux modèles distincts : le modèle relationnel et le modèle orienté objet. De plus, nous avons défini un modèle sémantique de documents multimédias plus précisément les documents scientifiques. En revanche, la diversité de mode de représentation de données multimédias implique pratiquement la variété des stratégies d'exploration de ce type de données. De ce fait, nous distinguons deux stratégies d'exploration de données multimédias : La recherche d'informations dans un corpus de documents multimédias et l'interrogation des bases de données multimédias. Sur la base de ces deux stratégies nous avons proposé deux approches distinctes, la première est orientée bases de données multimédias et la deuxième est orientée documents multimédias.

Dans le deuxième aspect, nous avons déterminé les différents problèmes liés à l'hétérogénéité sémantique des sources de données multimédias et nous avons focalisé notre travail aux travaux d'intégration de données, plus précisément la médiation sémantique à base d'ontologie pour assurer l'interrogation des bases de données multimédias et hétérogènes. Les travaux existants sont basés sur la médiation sémantique des bases de données relationnelles ne contenants pas des attributs de types complexe comme image, texte, audio et la vidéo. Notre première approche présente un travail original dans le sens où les bases de données à intégrer peuvent contenir des données multimédias qui sont définies par les types LOB. En ce qui concerne les documents multimédias, des travaux sont basés soient sur la personnalisation d'accès aux données ou sur l'indexation sémantique à base d'ontologie. Dans la deuxième approche proposée, nous présentons une nouvelle technique d'indexation de documents multimédias en utilisant le profil utilisateur. Ce dernier est utilisé comme élément de référence pour améliorer l'indexation de document plutôt que l'amélioration d'accès aux données car, la qualité du système de recherche d'informations ne dépend pas seulement à la bonne spécification des besoins d'utilisateur mais aussi à la qualité des descripteurs de données.

Dans le troisième aspect, l'ontologie joue un rôle crucial pour définir le sens de données et donc elle permet de traiter tous problèmes ou conflits sémantiques lors de l'exploration des sources de données multimédias. Nous utilisons les ontologies pour assurer l'intégration, la description et le traitement d'hétérogénéité sémantique de sources de données multimédias.

Nous présentons dans ce qui suit nos principales contributions pour le traitement du problème d'hétérogénéité sémantique pour l'exploration des sources de données multimédias.

1.1. Approche de médiation sémantique

Les données multimédias peuvent être représentées sous forme des n-uplets (enregistrements) stockées dans des bases de données (BD) ayant la capacité de représenter et de manipuler ce type de données, ces BD sont appelées bases de données multimédias (BDMM). La création de nos sources de données de domaine de risques alimentaires est faite par six concepteurs des BD qu'ils ont construit six BDMM: quatre conçues par le modèle relationnel et deux autres sont modélisées par le modèle orienté-objet. De ce fait, nous avons utilisé deux systèmes de gestion de bases de données (SGBD) : Oracle Database pour les BDMM à base du modèle relationnel et O2 dédié à la gestion des BDMM orientée-objet.

Notre contribution permet d'assurer une intégration sémantique des bases de données multimédias et hétérogènes. Elle s'appuie sur l'approche de médiation sémantique par hybridation et l'approche LAV (Local As View) pour assurer le mapping entre schéma global du médiateur et les schémas locaux des sources de données multimédias (BDMM). Le schéma global est défini par la création d'une ontologie partagée ONTARIS (ONTology of Alimentation RISks) dédié au domaine des risques alimentaires.

L'approche proposée vise à créer et à gérer automatiquement des ontologies virtuelles à partir des sources de données. Ces ontologies sont utilisées comme des vues sémantiques définissant à la fois la sémantique de données multimédias et les schémas locaux des BDMM à intégrer. Notre approche est fondée sur un processus d'appariement entre les éléments composants la requête utilisateur exprimée en termes du vocabulaire d'ONTARIS et ceux des ontologies virtuelles propres à chaque BDMM. L'avantage principal de ce processus est la capacité d'extraire toutes les relations sémantiques qui ne sont pas présentées dans les approches existantes fondées sur la traduction de requête du médiateur en des requêtes exprimées en termes de langage source. De plus, aucune réécriture de la requête initiale en termes de vues, ni la traduction des requêtes par les adaptateurs. La requête utilisateur a été transmise directement aux adaptateurs qui assurent son exécution via notre processus d'appariement.

Le processus d'appariement proposé prend en entrée à la fois la requête initiale soumise par l'utilisateur et l'ontologie virtuelle, et donne en sortie l'ensemble de correspondances répondant à cette requête. Ce processus se déroule en quatre phases successives: appariement concepts, appariement instances, appariement propriétés et appariement relations. Ces phases sont basées sur l'utilisation de WordNet pour désambiguïser le sens des mots et l'application de deux mesures de similarité Wu et Palmer et similarité Cosinus. De plus, les résultats obtenus dans chaque adaptateur ont été transmis à la couche médiateur qui sert à accomplir deux principales étapes : l'étape de prétraitement pour éliminer les redondances des réponses et l'étape de fusion pour assembler les réponses retenues de l'étape précédente afin de retourner à l'utilisateur des réponses normalisées comme si c'était une interrogation usuelle d'une seule base de données.

Sur le plan pratique, nous avons développé le système de médiation sémantique SAMER (SemAntic Mediation for alimEntation Risks) basé sur l'architecture en trois couches

(médiateur, adaptateurs et sources de données). De plus, nous avons mené deux grandes expérimentations :

- Evaluation des performances qui s'appuie sur le calcul de précision, rappel et F-mesure pour évaluer les performances du système en termes de qualité de réponses retournées. Les résultats montrent que SAMER capable d'intégrer toutes les données pertinentes et plus rarement de bruit (données intégrées non pertinentes);
- Comparaison qui se base sur deux études comparatives : une comparaison quantitative des mesures de similarité utilisées par rapport les autres mesures et une comparaison qualitative qui vise à montrer l'originalité et l'efficacité de notre approche par rapport les travaux similaires. Sur la base des résultats que nous avons obtenus, nous peuvent dire que le choix des mesures de similarité ainsi que l'approche de médiation proposée pour à la fois les BDMM relationnelle et orientées-objet représente une meilleure solution pour les données multimédias, tout cela nous a permis d'atteindre convenablement les objectifs de cette thèse.

Finalement, notre système est très évolutif dans le sens où il pourra être facilement enrichi par une nouvelle base de données et éventuellement un nouvel adaptateur.

1.2. Indexation personnalisée

La deuxième contribution présentée dans cette thèse s'inscrit dans le cadre du traitement du problème d'hétérogénéité sémantique pour la recherche dans un corpus de documents multimédias, plus précisément les documents scientifiques. L'approche proposée concerne l'indexation personnalisée à base du profil utilisateur. Cette approche s'inspire des travaux menés dans le domaine de la recherche d'informations personnalisée pour concevoir des index que l'on peut personnaliser à partir du modèle sémantique de document et celui du profil utilisateur.

L'indexation personnalisée de documents multimédias vise à enrichir les index de base des documents par des entités complémentaires en exploitant les concepts communs entre la représentation sémantique du document multimédia et celle du profil utilisateur.

Personnaliser l'indexation de documents multimédias et hétérogènes va permettre de réduire le problème d'hétérogénéité sémantique et donc d'explorer conformément le corpus de documents.

De plus, le modèle proposé pour l'indexation personnalisée de documents scientifiques et multimédias est applicable pour tous types de documents scientifiques (livres, mémoires, thèses, résultats d'analyse, …etc.), il suffit seulement d'utiliser l'ontologie de domaine approprie à la place de notre ontologie IROnto (Information Retrieval Ontology).

Le processus d'indexation personnalisée est donc avantageux pour l'utilisateur qui définit correctement son profil. Ce processus a été validé par l'outil Persodexing (Personalized indexing). De plus, une étude expérimentale a été établie en comparant notre processus d'indexation personnalisée avec une indexation classique. Les résultats obtenus

sont très encourageants en termes de qualité des index retournés et le temps écoulé pour indexer un document.

2. Perspectives

Outre les contributions présentées dans cette thèse, nous désirons dans les travaux futurs d'apporter des améliorations et des orientations dans les deux approches proposées :

1. ***Dans l'approche de médiation sémantique des BDMM :*** nous prévoyons de proposer une méthode d'enrichissement automatique de l'ontologie partagée ONTARIS à partir des sources de données hétérogènes afin d'être capable d'exprimer rigoureusement la requête d'utilisateur et d'explorer au mieux les bases de données multimédias. En ce qui concerne les données multimédias, nous voulons étudier d'autres types de média tels que la vidéo et le son en utilisant Oracle interMedia qui étend Oracle Database pour gérer les données multimédias par son principal package ORDSYS. En outre d'hétérogénéité des données existantes dans des sources de données, elles sont devenues récemment très volumineuses et elles sont produites plus rapidement. Ces caractéristiques nous ont conduits à orienter nos recherches vers l'intégration de big data, une première ébauche a été présentée dans [Aggoune, 16]. Nous avons présenté une architecture étendue du système de médiation SAMER pour les données de big data en se basant sur l'exploration des résumés qui simplifient la représentation de données à grand échelle. De ce fait, chaque adaptateur est muni d'un processus de résumé à base d'ontologie pour compresser les données volumineuses et donc facilite l'intégration et l'exploration de ces données.
2. ***Dans l'approche d'indexation personnalisée de documents multimédias :*** une première perspective que nous devons faire est d'utiliser l'outil d'indexation personnalisée Persodexing dans un système de recherche d'informations (SRI). Il serait intéressant de mettre en œuvre un SRI spécialement pour l'exploration de documents scientifiques et multimédias qui couvrent tous les domaines d'informatique. Ce système peut être vu comme un moteur de recherche de documents de domaine informatique. Par ailleurs, pour rendre l'outil Persodexing facilement utilisable, nous avons pensé à alléger la tâche de création et d'évolution du profil utilisateur qui joue un rôle très important dans l'indexation personnalisée de documents multimédias. Nous prévoyons de proposer à l'utilisateur une liste des termes suggérés obtenus à partir de son utilisation de l'outil et le contenu du son profil. L'utilisateur a le choix d'accepter ou de rejeter ces suggestions.

BIBLIOGRAPHIE

[Abiteboul, 02] Abiteboul, S., Cluet, S., Ferran, G. et Rousset, M. C. (2002) 'The xyleme project', Computer Networks, Vol. 39, No. 3, pp. 225-238.

[Aggoune, 17] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2017) 'Mediation system for dealing with semantic problems in databases', International Journal of Data Mining, Modelling and Management, IJDMMM, ISSN : 1759-1163, Inderscience Publishers Ltd, (In press).

[Aggoune, 16] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2016) 'Big Data Integration: A Semantic Mediation Architecture Using Summary', International Conference on Advanced Technologies for Signal and Image Processing, ATSIP'2016, Monastir, Tunisie, IEEE.

[Aggoune, 15a] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2015) 'A Extended Semantic Indexing Model for Heterogeneous Corpus: Case Study in Scientific Documents', Colloque TASSILI Système Conjoint de Compression et d'Indexation Basé-Objet pour la Vidéo, Tassili-SCCIBOV'2015, Sidi bel Abbes, Algérie.

[Aggoune, 15b] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2015) 'A New semantic proximity measure for fuzzy query optimization in relational databases', International Conference on Pattern Analysis and Intelligence Systems, PAIS'2015, Tebessa, Algérie, IEEE, (*First Best Paper*).

[Aggoune, 14a] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2014) 'Personnalized indexing for heterogeneous multimedia data', 4th International Symposium ISKO-Maghreb'2014 Concepts and Tools for Knowledge Management (KM), CERIST, Alger, Algérie, IEEE.

[Aggoune, 14b] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2014) 'Personnalisation d'accès aux sources de données hétérogènes pour l'organisation des grands systèmes d'information d'entreprise', International Conference on Information Technology for Organization Development, IT4OD'2014, Tebessa, Algérie, IEEE.

[Aggoune, 13a] Aggoune, A. Bouramoul, A., Kholladi, M.K. et Doan, B.L. (2013) 'Geometric Transformation of User Queries in Information Retrieval on the Web', International Arab Conference on Information Technology, ACIT'2013, Soudan.

[Aggoune, 13b] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2013) 'Approche à base de requêtes géométriques pour l'amélioration de recherche d'information sur le web', 2ème journées doctorales du laboratoire de modélisation et d'implémentation des systèmes complexes, JDMISC'2013, Constantine, Algérie.

[Aggoune, 13c] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2013) 'Improving web search by using geometric queries: A Novel approach', Journée doctorale du laboratoire des sciences et technologies de l'information et de la communication, JSTIC'2013, Guelma, Algérie.

[Aggoune, 12a] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2012) 'Problème de l'hétérogénéité sémantique de documents multimédias', Deuxième Journée doctorale en informatique, JDI'2012, Guelma, Algérie.

[Aggoune, 12b] Aggoune, A. Bouramoul, A. et Kholladi, M.K. (2012) 'Approximate Flexible Queries Using Hausdorff Distance', Second International Symposium on Modelling and Implementation of Complex Systems, MISC'2012, Constantine, Algérie.

[Ahn, 12] Ahn, B. T. et Kim, M. S. (2012) 'Embedded Multimedia DBMS based on Mpeg-7 in Mobile Environment', International Information Institute (Tokyo), Information, Vol. 15, No. 11, 5041.

[Albertoni, 11] Albertoni, R., Camossi, E., De Martino, M., Giannini, F. et Monti, M. (2011) 'Context dependent semantic granularity', International Journal of Data Mining, Modelling and Management, Vol. 3, No.2, pp.189–215.

[Allen, 09] Allen, G., Bryla, B. et Kuhn, D. (2009) 'LOBs', In Oracle SQL Recipes, pp. 383-400, Apress.

[Ali, 09] Ali, M.G. (2009) 'Object-Oriented Approach for Integration of Heterogeneous

Databases in a Multidatabase System and Local Schemas Modifications Propagation', arXiv preprint arXiv:0912.0603.

[Amann, 93] Amann, B., Christophides, V. et Scholl, M. (1993) 'HyperPATH/O2: Integrating hypermedia systems with object-oriented database systems', In International Conference on Database and Expert Systems Applications, pp. 709-720, Springer Berlin Heidelberg.

[Amato, 04] Amato, G., Gennaro, C., Rabitti, F. et Savino, P. (2004) 'Milos: A multimedia content management system for digital library applications', In International Conference on Theory and Practice of Digital Libraries, pp. 14-25, Springer Berlin Heidelberg.

[Amous, 02] Amous, I., Jedidi, A. et Sèdes, F. (2002) 'A contribution to multimedia document modeling and organizing', In International Conference on Object-Oriented Information Systems, pp. 434-444, Springer Berlin Heidelberg.

[Anderton, 97] Anderton, R. L. (1997) 'Pixel artifact/blemish filter for use in CCD video camera', U.S. Patent, No 5,619,261.

[Angus, 13] Angus, D., Rintel, S. et Wiles, J. (2013) 'Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis', International Journal of Social Research Methodology, Vol. 16, No. 3, pp. 261-267.

[Antoniou, 04] Antoniou, G. et Van Harmelen, F. (2004) 'Web ontology language: Owl', In Handbook on ontologies, pp. 67-92, Springer Berlin Heidelberg.

[Arens, 93] Arens, Y. et Knoblock, C. (1993) 'Sims: Retrieving and integrating information from multiple sources', In ACM SIGMOD Record, Vol. 22, No. 2, pp. 562-563, ACM.

[Aversano, 02] Aversano, L., Canfora, G., De Lucia, A. et Stefanucci, S. (2002) 'Understanding SQL through iconic interfaces', In Proceedings of International conference of Computer Software and Applications Conference, COMPSAC 2002, pp. 703-708, IEEE.

[Baader, 99] Baader, F., Küsters, R., Borgida, A. et McGuinness, D. L. (1999) 'Matching in description logics', Journal of Logic and Computation, Vol. 9, No. 3, pp.411-447.

[Baccini, 12] Baccini, A., Déjean, S., Lafage, L. et Mothe, J. (2012) 'How many performance measures to evaluate Information Retrieval Systems?', Knowledge and Information Systems, Vol. 30, No. 3, pp.693.

[Bachimont, 00] Bachimont B. (2000) 'Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances', Ingénierie des connaissances : évolution Récentes et nouveaux défis Paris: Eyrolles, pp.305-323

[Bancilhon, 88] Bancilhon, F., Barbedette, G., Benzaken, V., Delobel, C., Gamerman, S., Lécluse, C. et al. (1988) 'The design and implementation of O2, an object-oriented database system', In International Workshop on Object-Oriented Database Systems, pp. 1-22, Springer Berlin Heidelberg.

[Bellatreche, 13] Bellatreche, L., Khouri, S. et Berkani, N. (2013) 'Semantic data warehouse design: From ETL to deployment à la carte', In International Conference on Database Systems for Advanced Applications, pp. 64-83, Springer Berlin Heidelberg.

[Bellatreche, 03] Bellatreche, L., Pierra, G., Xuan, D. N. et Hondjack, D. (2003) 'An Automated Information Integration Technique using an Ontology-based Database Approach', In Proceeding of 10th ISPE International Conference on Concurrent Engineering: Research and Applications, Special Track on Data Integration in Engineering, pp 217-223, 2003.

[Beneventano, 13] Beneventano, D., Gennaro, C., Bergamaschi, S. et Rabitti, F. (2013) 'A mediator-based approach for integrating heterogeneous multimedia sources', Multimedia tools and applications, Vol. 62, No. 2, pp. 427-450.

[Beneventano, 01] Beneventano, D., Bergamaschi, S., Guerra, F. et Vincini, M. (2001) 'The momis

approach to information integration', pp. 194-198.

[Berners-Lee, 01] Berners-Lee, T., Hendler J. et Lassila, O. (2001) 'The semantic web', Scientific american, Vol. 284, No.5, pp. 28-37.

[Biebow, 99] Biebow, B., Szulman, S. et Clément, A. J. (1999) 'TERMINAE: A linguistics-based tool for the building of a domain ontology', In International Conference on Knowledge Engineering and Knowledge Management, pp. 49-66, Springer Berlin Heidelberg.

[Bloehdorn, 05] Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V. et al (2005) 'Semantic annotation of images and videos for multimedia analysis', In European Semantic Web Conference, pp. 592-607, Springer Berlin Heidelberg.

[Boisot, 04] Boisot, M. et Canals, A. (2004) 'Data, information and knowledge: have we got it right?', Journal of evolutionary economics, Vol. 14, No.1, pp.43-67.

[Bond, 12] Bond, F. et Paik, K. (2012) 'A survey of wordnets and their licenses', Small, Vol. 8, No. 4, pp. 1-8.

[Borst, 97] Borst, W. N. (1997) 'Construction of engineering ontologies for knowledge sharing and reuse'. Thèse de doctorat en informatique, 243 pages, université de TWENTE, Enschede, Pays-Bas.

[Bouchou, 14] Bouchou, B. et Niang, C. (2014) 'Semantic mediator querying', In Proceedings of the 18th International Database Engineering & Applications Symposium, pp. 29-38, ACM.

[Bouguila, 07] Bouguila, N. et Ziou, D. (2007) 'Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization', Journal of Visual Communication and Image Representation, Vol. 18, No. 4, pp. 295-309.

[Boulçane, 08] Boulçane, F. et Boufaida, M. (2008) 'A Hybrid Data Integration Approach based on specialized mediators', In proceedings of the International Review on Computers and Software, Vol.3 No.5, pp. 554-563.

[Bouramoul, 10] Bouramoul, A., Kholladi, M.K. et Doan, B.L. (2010) 'PRESY: A context based query reformulation tool for information retrieval on the web', Journal of Computer Science, Vol. 6, No. 4, pp. 470-477.

[Bouzeghoub, 05] Bouzeghoub, M. et Kostadinov, D. (2005) 'Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils', CORIA, Vol. 5, pp. 201-218.

[Boyd, 02] Boyd, M., McBrien, P. et Tong, N. (2002) 'The automed schema integration repository', In British National Conference on Databases, pp. 42-45, Springer Berlin Heidelberg.

[Bozsak, 02] Bozsak, E., Ehrig, M., Handschuh, S., Hotho, A. et al. (2002) 'KAON - Towards a large scale Semantic Web', In Proceedings of the 3rd International Conference on E-Commerce and Web Technologies (ECWeb'2002), Vol. 2455, pp 304-313.

[Brachman, 85] Brachman, R. J. et Schmolze, J. G. (1985) 'An overview of the KL-ONE knowledge representation system', Cognitive science, Vol. 9, No. 2, pp.171-216.

[Braga, 03] Braga, D. et Campi, A. (2003) 'A graphical environment to query xml data with xquery', In Proceedings of the Fourth International Conference of Web Information Systems Engineering, WISE 2003, pp. 31-40, IEEE.

[Buccella, 05] Buccella, A., Cechich, A. and Brisaboa, N. R. (2005) 'Ontology-based data integration methods: a framework for comparison', Journal of Intelligent and Cooperative Information Systems, Vol. 2, No.2, pp.127–158.

[Buchanan, 92] Buchanan, M. C. et Zellweger, P. T. (1992) 'Specifying temporal behavior in hypermedia documents', In Proceedings of the ACM conference on Hypertext, pp. 262-271, ACM.

[Bulterman, 02] Bulterman, D. C. A. (2002) 'SMIL 2.0. 2. Examples and comparisons', IEEE

MultiMedia, Vol. 9, No. 1, pp. 74-84.

[Buil-Aranda, 13] Buil-Aranda, C., Arenas, M., Corcho, O. et Polleres, A. (2013) 'Federating queries in SPARQL 1.1: Syntax, semantics and evaluation', Web Semantics: Science, Services and Agents on the World Wide Web, Vol.18, No.1, pp.1-17.

[Calì, 13] Calì, A., Calvanese, D., De Giacomo, G. et Lenzerini, M. (2013) 'Data integration under integrity constraints', In Seminal Contributions to Information Systems Engineering, pp. 335-352, Springer Berlin Heidelberg.

[Carey, 95] Carey, M. J., Haas, L. M., Schwarz, P. M., Arya, M., et al. (1995), 'Towards heterogeneous multimedia information systems: The Garlic approach', In Proceedings of Fifth International Workshop on Research Issues in Data Engineering: Distributed Object Management, pp. 124-131, IEEE.

[Chang, 80] Chang, N. S. et Fu, K. S. (1980) 'Query-by-pictorial-example', IEEE Transactions on Software Engineering, No. 6, pp.519-524.

[Chawathe, 94] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. et Widom, J. (1994) 'The TSIMMIS project: Integration of heterogenous information sources'.

[Cheatham, 13] Cheatham, M. et Hitzler, P. (2013) 'String similarity metrics for ontology alignment', In International Semantic Web Conference, pp. 294-309, Springer Berlin Heidelberg.

[Christodoulakis, 86] Christodoulakis, S., Ho, F. et Theodoridou, M. (1986) 'The multimedia object presentation manager of MINOS: a symmetric approach', In ACM SIGMOD Record, Vol. 15, No. 2, pp. 295-310.

[Chupeau, 03] Chupeau, P. et Tazine, N. (2003) 'ANNAPURNA : Annotation Automatique d'images pour la Recherche et la Navigation', Rapport, Modélisation et Recherche d'Information Multimédia, pp.1-7.

[Clark, 02] Clark, P. et Mirmehdi, M. (2002) 'Recognising text in real scenes', International Journal on Document Analysis and Recognition, Vol. 4, No. 4, pp.243-257.

[Codd, 70] Codd, E. F. (1970) 'A relational model of data for large shared data banks', Communications of the ACM, Vol. 13, No. 6, pp. 377-387.

[Das, 15] Das, D., Yan, J., Zait, M., Valluri, S. R., Vyas, N., Krishnamachari, R. et al. (2015) 'Query optimization in Oracle 12c database in-memory', Proceedings of the VLDB Endowment, Vol. 8, No. 12, pp.1770-1781.

[Dasan, 98] Dasan, V.S. (1998) 'Personalized information retrieval using user-defined profile', U.S. Patent, No 5,761,662.

[Dean, 04] Dean, M., Schreiber, G., Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks et al. (2004) 'OWL web ontology language reference', W3C Recommendation, https://www.w3.org/TR/2004/REC-owl-ref-20040210/

[De Giacomo, 12] De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rosati, R., Ruzzi, M. and Savo, D.F. (2012) 'MASTRO: a reasoner for effective ontology-based data access', in ORE 2012: Proceeding of the first International Workshop on OWL Reasoner Evaluation, Manchester, Vol. 858, pp.1–11.

[Derbas, 14] Derbas, N. et Quénot, G. (2014) 'Mots audio-visuels joints pour la détection de scènes violentes dans les vidéos', CORIA, pp. 63-77.

[Dessloch, 97] Dessloch, S. et Mattos, N. (1997) 'Integrating SQL databases with content-specific search engines', In VLDB, Vol. 97, pp. 528-537.

[De Valk, 15] de Valk, H. et Salvat, G. (2015) 'Alimentation et risques infectieux: enjeux et stratégies pour limiter l'impact sur la santé', Les Tribunes de la santé, No. 4, pp.61-68.

[Delmal, 00] Delmal, P. (2000) 'SQL2-SQL3: applications à Oracle', Livre, édition 3, De Boeck Supérieur, 509 pages.

[Djema, 07] Djema, L., Boumghar, F. et Debiane, S. (2007) 'L'imagerie Médicale Dans une

Base De Données Distribuée Multimédia Sous Oracle 9i', In Proceeding of the 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March, pp. 25-29, Tunisie.

[Drame, 14] Drame k. (2014) 'Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical', Thèse de doctorat en informatique, 187 pages, université de Bordeaux, France.

[Egghe, 08] Egghe, L. (2008) 'The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations', Information Processing & Management, Vol. 44, No. 2, pp.856-876.

[Ehring, 07] Ehring M. (2007) 'Ontology Alignment: Bridging the Semantic Gap: Semantic Web and Beyond', Springer-Verlag US, pp. 61-96.

[Ehrig, 04] Ehrig, M and Sure, Y. (2004) 'Ontology mapping–an integrated approach', in ESWS 2004: Proceeding of the 1st European Semantic Web Symposium, Heraklion, Greece, pp.76–91.

[Elbyed, 09] Elbyed, A. (2009) 'ROMIE, une approche d'alignement d'ontologies à base d'instances', Thèse de doctorat, Institut National des Télécommunication, 182 pages.

[Elmasri, 00] Elmasri, R. et Navathe, S. (2000) 'Fundamentals of database systems', Addison-Wesley.

[Eidenberger, 04] Eidenberger, H. (2004) 'Statistical analysis of content-based MPEG-7 descriptors for image retrieval', Multimedia Systems, Vol. 10, No.2, pp.84-97.

[Eisenberg, 99] Eisenberg, A. et Melton, J. (1999) 'SQL: 1999, formerly known as SQL3', ACM Sigmod record, Vol. 28, No. 1, pp.131-138.

[Esuli, 07] Esuli, A. et Sebastiani, F. (2007) 'SENTIWORDNET: A high-coverage lexical resource for opinion mining', Evaluation, pp.1-26.

[Euzenat, 07] Euzenat, J. et Shvaiko, P. (2007) 'Ontology matching', Livre, Vol. 18, 511 pages, Heidelberg: Springer.

[Euzenat, 04] Euzenat, J., Loup, D., Touzani, M. et Valtchev, P. (2004) 'Ontology alignment with OLA', In Proc. 3rd ISWC2004 workshop on Evaluation of Ontology-based tools, EON, pp. 59-68.

[Fellah, 08] Fellah, A., Malki, M. et Zahaf, A. (2008) 'Alignement des ontologies: utilisation de WordNet et une nouvelle mesure structurelle', In Conférence en Recherche d'Information et Applications, CORIA, pp. 401-408.

[Feinberg, 06] Feinberg, M., Bertail, P., Tressou-Cosmao, J. et Verger, P. (2006) 'Analyse des risques alimentaires', Tec et Doc, 416 pages.

[Feng, 13] Feng, D., Siu, W. C. et Zhang, H. J. (2013) 'Multimedia information retrieval and management: Technological fundamentals and applications', Livre, Springer Science & Business Media.

[Ferreira, 01] Ferreira, J. et da Silva, A. R. (2001) 'MySDI: A Generic Architecture to Develop SDI Personalised Services', In ICEIS, Vol. 1, pp. 262-270.

[Fernández-López, 97] Fernández-López, M., Gómez-Pérez, A. et Juristo, N. (1997) 'Methontology: from ontological art towards ontological engineering', In AAAI-97, Spring Symposium Series, Université de Stanford.

[Finlayson, 14] Finlayson, M. A. (2014) 'Java libraries for accessing the princeton wordnet: Comparison and evaluation', In Proceedings of the 7th Global Wordnet Conference, Tartu, Estonia.

[Friedman, 99] Friedman, M., Levy, A. et Millstein, T. (1999) 'Navigational plans for data integration', In Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications, pp. 67–73.

[Gao, 13] Gao, Y., Wang, M., Zha, Z. J., Shen, J., Li, X. et Wu, X. (2013) 'Visual-textual joint relevance learning for tag-based social image search', IEEE Transactions on Image Processing, Vol. 22, No. 1, pp.363-376.

[Ganchev, 05] Ganchev, T., Fakotakis, N. et Kokkinakis, G. (2005) 'Comparative evaluation of various MFCC implementations on the speaker verification task', In Proceedings of the SPECOM, Vol. 1, pp. 191-194.

[Gardarin, 03] Gardarin, G. (2003) 'Bases de données', Livre, 5ème édition, editions eyrolles, 788 pages.

[Gatziu, 98] Gatziu, S., Vavouras, A. et Dittrich, K. R. (1998) 'Sirius: An approach for data warehouse refreshment', LIVRE, Universität Zürich. Institut für Informatik.

[Gauch, 07] Gauch, S., Speretta, M., Chandramouli, A. et Micarelli, A. (2007) 'User profiles for personalized information access', In The adaptive web, pp. 54-89, Springer Berlin Heidelberg.

[Gavini, 11] Gavini, M. (2011) 'Oracle : exploitation des bases de données en environnement de production sous Unix', livre, ENI edition, 511 pages.

[Genesereth, 97] Genesereth, M. R., Keller, A. M. et Duschka, O. M. (1997) 'Infomaster: An information integration system', In ACM SIGMOD Record, Vol. 26, No. 2, pp. 539-542, ACM.

[Gennaro, 11] Gennaro, C., Lenzi, R., Mandreoli, F., Martoglia, R., Mordacchini, M., Penzo, W. et Sassatelli, S. (2011) 'A unified multimedia and semantic perspective for data retrieval in the semantic web', Information Systems, Vol. 36, No. 2, pp.174-191.

[Gesmundo, 12] Gesmundo, A. et Samardžić, T. (2012) 'Lemmatisation as a tagging task', In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Vol. 2, pp. 368-372, Association for Computational Linguistics.

[Getman, 14] Getman, A. P. et Karasiuk, V. V. (2014) 'A crowdsourcing approach to building a legal ontology from text', Artificial Intelligence and Law, Vol. 22, No.3, pp.313-335.

[Ghosh, 16] M. El Ghosh, H. Naja, H. Abdulrab, and M. Khalil (2016) 'Towards a Middle-out Approach for Building Legal Domain Reference Ontology', International Journal of Knowledge Engineering, Vol. 2, No. 3, pp.109-114

[Goh, 99] Goh, C. H., Bressan, S., Madnick, S. et Siegel, M. (1999) 'Context interchange: New features and formalisms for the intelligent integration of information', ACM Transactions on Information Systems (TOIS), Vol. 17, No. 3, pp.270-293.

[Gómez-Pérez, 02] Gómez-Pérez, A. et Corcho, O. (2002) 'Ontology languages for the semantic web', IEEE Intelligent systems, Vol. 17, No. 1, pp.54-60.

[Gómez-Pérez, 99] Gómez-Pérez A. (1999) 'Ontological Engineering: A state of the art', Expert Update: Knowledge Based Systems and Applied Artificial Intelligence, Vol. 2, No. 3, pp.33-43.

[Gray, 97] Gray, P.M.D., Preece, A.D., Fiddian, N.J. et al. (1997) 'KRAFT: knowledge fusion from distributed databases and knowledge bases', in DEXA 1997: Database and Expert System Applications Workshop, Toulouse, France, pp.682–691.

[Gruninger, 95] Gruninger, M. et Fox, M. S. (1995) 'The role of competency questions in enterprise engineering', In Benchmarking, Theory and practice, pp. 22-31, Springer US.

[Gruber, 93] Gruber, T. R. (1993) 'A translation approach to portable ontologies', Knowledge Acquisition, Vol. 5, No. 2, pp.199-229.

[Guarino, 09] Guarino, N., Oberle, D. et Staab, S. (2009) 'What is an Ontology?', In Handbook on ontologies, pp.1-17, Springer Berlin Heidelberg.

[Guarino, 97a] Guarino N. (1997) 'Some organizing principles for a unified top-level ontology', AAAI Spring Symposium on Ontological Engineering, pp.57-63.

[Guarino, 97b] Guarino N. (1997) 'Understanding, building and using ontologies', International

Journal of Human-Computer Studies, Vol.46, No. 2-3, pp.293-310.

[Gudivada, 95] Gudivada, V. N. et Raghavan, V. V. (1995) 'Content based image retrieval systems', Computer, Vol. 28, No. 9, pp.18-22.

[Guo, 14] Guo, K., Ma, J. et Duan, G. (2014) 'Dhsr: a novel semantic retrieval approach for ubiquitous multimedia', Wireless Personal Communications, Vol. 76, No. 4, pp.779-793.

[Halevy, 01] Halevy, A.Y. (2001) 'Answering queries using views: a survey', International Journal on Very Large Data Bases, Vol. 10, No. 4, pp.270–294.

[Hamadi, 15] Hamadi, A., Mulhem, P. et Quénot, G. (2015) 'Extended conceptual feedback for semantic multimedia indexing', Multimedia Tools and Applications, Vol. 74, No. 4, pp. 1225-1248.

[Heflin, 98] Heflin, J., Hendler, J. et Luke, S. (1998) 'Reading between the lines: Using SHOE to discover implicit knowledge from the Web', In AAAI-98 Workshop on AI and Information Integration, Vol. 297.

[Heimbigner, 85] Heimbigner, D. et McLeod, D. (1985) 'A federated architecture for information management', ACM Transactions on Information Systems, TOIS, Vol. 3, No. 3, pp.253-278.

[Hepp, 07] Hepp, M. et de Bruijn, J. (2007) 'GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies', In European Semantic Web Conference, pp. 129-144, Springer Berlin Heidelberg.

[Hoepner, 92] Hoepner, P. (1992) 'Synchronizing the presentation of multimedia objects', Computer Communications, Vol. 15, No. 9, pp.557-564.

[Horrocks, 02] Horrocks, I. (2002) 'DAML+OIL: A Description Logic for the Semantic Web', IEEE Data Eng. Bull., Vol. 25, No. 1, pp. 4-9.

[Inmon, 00] Inmon, W.H. (2000) 'Building the data warehouse', livre, printed by united states of America, 407 pages.

[Inmon, 93] Inmon, W.H. et Chuck Kelley. (1993) 'RDB/VMS: Developing the Data Warehouse', Livre. Boston. QED Pub Group.

[Ives, 04] Ives, Z. G., Halevy, A. Y., Mork, P. et Tatarinov, I. (2004) 'Piazza: mediation and integration infrastructure for semantic web data', Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 1, No. 2, pp.155-175.

[Jansen, 13] Jansen, J., Cesar, P. et Bulterman, D. (2013) 'Multimedia document synchronization in a distributed social context', In Proceedings of the 2013 ACM symposium on Document engineering, pp. 273-276.

[Jain, 13] Jain, V. et Singh, M. (2013) 'Ontology based information retrieval in semantic web: A survey', International Journal of Information Technology and Computer Science, IJITCS, Vol. 5, No. 10, pp.62.

[Jarke, 13] Jarke, M., Jeusfeld, M. A., Quix, C. J., Vassiliadis, P. et Vassiliou, Y. (2013) 'Data warehouse architecture and quality: impact and open challenges', In Seminal Contributions to Information Systems Engineering, pp. 183-189, Springer Berlin Heidelberg.

[Jarke, 97] Jarke, M. et Vassiliou, Y. (1997) 'Data Warehouse Quality: A Review of the DWQ Project', In 2nd Conference on Information Quality (IQ), pp. 299-313.

[Jean-Mary, 09] Jean-Mary, Y. R., Shironoshita, E. P. et Kabuka, M. R. (2009) 'Ontology matching with semantic verification', Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 7, No. 3, pp. 235-251.

[Jedidi, 05] Jedidi A. (2005) 'Modélisation générique de documents multimédia par des métadonnées : mécanismes d'annotation et d'interrogation', thèse de doctorat, Université Paul Sabatier-Toulouse III.

[Jiang, 97] Jiang, J.J et Conrath, D.W. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy', in ROCLING 1997: Proceedings of the International Conference on Research in Computational Linguistics, pp.507–514.

[Jorgensen, 13] Jorgensen, C. (2013) 'The MPEG-7 Initiative for Multimedia Content Description', In Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.

[Kalyanpur, 06] Kalyanpur, A., Parsia, B., Sirin, E., Grau, B. C. et Hendler, J. (2006) 'Swoop: A web ontology editing browser', Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 4, No. 2, pp.144–153.

[Kalyanpur, 05] Kalyanpur, A., Parsia, B. et Hendler, J. (2005) 'A tool for working with web ontologies', International Journal on Semantic Web and Information Systems, IJSWIS, Vol. 1, No. 1, pp. 36-49.

[Kent, 99] Kent, R.E. (1999) 'Conceptual Knowledge Markup Language: The Central Core', In Proceedings of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management.

[Khouri, 12] Khouri, S., Ilyès, B., Ladjel, B., Eric, S., Stéphane, J. et Michael, B. (2012) 'Ontology-based structured web data warehouses for sustainable interoperability: requirement modeling, design methodology and tool', Journal of Computers in Industry, Vol. 63 No.8, pp.799–812.

[Knoll, 98] Knoll, A., Altenschmidt, C., Biskup, J., Blüthgen, H. M., Glöckner, et al. (1998), 'An integrated approach to semantic evaluation and content-based retrieval of multimedia documents', In International Conference on Theory and Practice of Digital Libraries, pp. 409-428, Springer Berlin Heidelberg.

[Knublauch, 04] Knublauch, H., Fergerson, R. W., Noy, N. F. et Musen, M. A. (2004) 'The Protégé OWL plugin: An open development environment for semantic web applications', In International Semantic Web Conference, pp. 229-243, Springer Berlin Heidelberg.

[Kolte, 08] Kolte, S. G. et Bhirud, S. G. (2008) 'Word sense disambiguation using wordnet domains', In International Conference on Emerging Trends in Engineering and Technology, ICETET'08, pp. 1187-1191, IEEE.

[Kozaki, 07] Kozaki, K., Sunagawa, E., Kitamura, Y. et Mizoguchi, R. (2007) 'A framework for cooperative ontology construction based on dependency management of modules', In Proceedings of the First International Conference on Emergent Semantics and Ontology Evolution, pp. 33-44, CEUR-WS. org.

[Krivine, 09] Krivine, S., Nobécourt, J., Soualmia, L., Cerbah, F. et Duclos, C. (2009) 'Construction automatique d'ontologie à partir de bases de données relationnelles: application au médicament dans le domaine de la pharmacovigilance', Actes des 20es Journées Francophones d'Ingénierie des Connaissances, IC' 2009, pp. 1-12

[Kustanowitz, 05] Kustanowitz, J. et Shneiderman, B. (2005) 'Motivating annotation for personal digital photo libraries: Lowering barriers while raising incentives', Univ. of Maryland Technical Report HCIL-2004, pp.18.

[Labio, 97] Labio, W. J., Zhuge, Y., Wiener, J. L., Gupta, H., Garcia-Molina, H. et Widom, J. (1997) 'The WHIPS prototype for data warehouse creation and maintenance', In ACM SIGMOD Record, Vol. 26, No. 2, pp. 557-559, ACM.

[Laborie, 08] Laborie, S. (2008) 'Adaptation sémantique de documents multimédia, thèse de doctorat, Université Joseph-Fourier-Grenoble I.

[Lakshman, 03] Lakshman, B. (2003) 'Floating in Java', In Oracle 9i PL/SQL: A Developer's Guide, pp. 509-554, Apress.

[Lamel, 08] Lamel, L. et Gauvain, J. L. (2008) 'Speech processing for audio indexing', Advances in Natural Language Processing, Springer Berlin Heidelberg, pp. 4-15.

[Lassila, 98] Lassila, O., Swick, R.R., Wide, W. et Consortium, W. (1998) 'Resource Description Framework (RDF) Model and Syntax Specification',

https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[Lazaridis, 13] Lazaridis, M., Axenopoulos, A., Rafailidis, D. et Daras, P. (2013) 'Multimedia search and retrieval using multimodal annotation propagation and indexing techniques', Signal Processing: Image Communication, Vol. 28, No. 4, pp. 351-367.

[Lenat, 89] Lenat, D.B. et Guha, R.V. (1989) 'Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project', 1st edition. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Lenzerini, 02] Lenzerini, M. (2002) 'Data integration: A theoretical perspective', In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 233-246, ACM.

[Levy, 96] Levy, A., Rajaraman, A. et Ordille, J. (1996) 'Querying heterogeneous information sources using source descriptions', In Proceedings of the Twenty-second International Conference on Very Large Data Bases (VLDB'96), pp. 251–262, Mumbai, India.

[Li, 13] Li, B. et Han, L. (2013) 'Distance weighted cosine similarity measure for text classification', In International Conference on Intelligent Data Engineering and Automated Learning, pp. 611-618, Springer Berlin Heidelberg.

[Li, 10] Li, L. J., Su, H., Fei-Fei, L. et Xing, E. P. (2010) 'Object bank: A high-level image representation for scene classification & semantic feature sparsification', In Advances in neural information processing systems, pp.1378-1386.

[Li, 07] Li, Q., Shi, Z. et Luo, S. (2007) 'Image retrieval based on fuzzy color semantics', Fuzzy Systems Conference, FUZZ-IEEE, IEEE International, pp. 1-5.

[Liu, 07] Liu, Y., Zhang, D., Lu, G. et Ma, W. Y. (2007) 'A survey of content-based image retrieval with high-level semantics', Pattern recognition, Vol. 40, No. 1, pp. 262-282.

[Looman, 60] Looman, J and Campbell, J.B. (1960) 'Adaptation of Sorensen's K for estimating unit affinities in prairie vegetation', International Journal of Ecology, Vol. 41, No. 3, pp.409–416.

[Lu, 15] Lu, C., Liu, M. et Wu, Z. (2015) 'SVQL: A SQL Extended Query Language for Video Databases', International Journal of Database Theory and Application, Vol. 8, No. 3, pp.235-248.

[Litwin, 89] Litwin, W., Abdellatif, A., Zeroual, A., Nicolas, B. et Vigier, P. (1989) 'MSQL: A multidatabase language', Information sciences, Vol. 49, No. 1, pp.59-101.

[Litwin, 85] Litwin, W. (1985) 'An overview of the multidatabase system MRDSM', In Proceedings of the 1985 ACM annual conference on The range of computing: mid-80's perspective: mid-80's perspective, pp. 524-533, ACM.

[Maedche, 03] Maedche, A., Motik, B., Stojanovic, L., Studer, R. et Volz, R. (2003) 'An infrastructure for searching, reusing and evolving distributed ontologies', In Proceedings of the 12th international conference on World Wide Web, pp. 439-448, ACM

[Maedche, 00] Maedche, A. et Staab, S. (2000) 'The text-to-onto ontology learning environment', In Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures, Vol. 38.

[Manjunath, 02] Manjunath, B. S., Salembier, P. et Sikora, T. (2002) 'Introduction to MPEG-7: multimedia content description interface', vol. 1, John Wiley & Sons.

[Marano, 13] Marano, F. et Guadagnini, R. (2013) 'Les nanoparticules dans l'alimentation: quels risques pour le consommateur?', Cahiers de Nutrition et de Diététique, Vol. 48, No. 3, pp. 142-150

[Maredj, 13] Maredj, A. et Tonkin, N. (2013) 'Gestion du recouvrement spatial dans les documents multimédia : Approche et évaluation', Revue TSI: Technique et science informatique, Hermès Lavoisier, Vol. 27, No.1, pp. 29-502.

[Maron, 60] Maron, M. E. et Kuhns, J. L. (1960) 'On relevance, probabilistic indexing and information retrieval', Journal of the ACM (JACM), Vol.7, No. 3, pp.216-244.

[Mattos, 99] Mattos, N. M., Darwen, H., Cotton, P., Pistor, P., Kulkarni, K., Dessloch, S. et Zeidenstein, K. (1999) 'Sql99, sql/mm, and sqlj: An overview of the sql standards', IBM Database Common Technology.

[Mbaiossoum, 12] Mbaiossoum, B., Khouri, S., Bellatreche, L., Jean, S., et Baron, M. (2012) 'Etude Comparative des Systèmes de Bases de Données à base Ontologiques', In INFORSID, pp. 379-394.

[Mbarki, 08] Mbarki M. (2008) 'Gestion de l'hétérogénéité documentaire: le cas d'un entrepôt de documents multimédia', Thèse de doctorat, Université de Toulouse III– Paul SABATIER.

[Mbarki, 07] Mbarki M., Soulé-Dupuy C. et Vallés-Parlangeau N. (2007) 'A Document Repository Architecture for Heterogeneous Business Information Management', International Conference on Enterprise Information Systems, ICEIS, INSTICC Press, pp. 192-198.

[McBride, 04] McBride, B. (2004) 'The resource description framework (RDF) and its vocabulary description language RDFS', In Handbook on ontologies, pp. 51-65, Springer Berlin Heidelberg.

[McGuinness, 04] McGuinness, D. L. et Van Harmelen, F. (2004) 'OWL web ontology language overview', W3C recommendation, Vol. 10, No. 10, 1-22.

[Melton, 01] Melton, J. et Eisenberg, A. (2001) 'SQL multimedia and application packages (SQL/MM)', ACM Sigmod Record, Vol. 30, No. 4, pp.97-102.

[Mena, 00] Mena, E., Illarramendi, A., Kashyap, V. et Sheth, A.P. (2000) 'OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies', International Journal of Distributed and parallel Databases, Vol. 8, No. 2, pp.223–271.

[Menon, 05] Menon, R. M. (2005) 'Using LOBs and BFILEs', Expert Oracle JDBC Programming, pp. 447-493.

[Michard, 98] Michard A. (1998) 'XML langage et applications', Livre, Edition Eyrolles.

[Midouni, 16] Midouni, S. A. D., Youssef, A. et Azeddine, C. (2016) 'Multimedia Data Retrieving based on SOA Architecture', Journal of Web Engineering, Vol. 15, No. 3et 4, pp. 339-360.

[Miller, 90] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. et Miller, K. J. (1990) 'Introduction to WordNet: An on-line lexical database', International journal of lexicography, Vol. 3, No. 4, pp. 235-244.

[Miller, 95] Miller, G. A. (1995) 'WordNet: a lexical database for English', Communications of the ACM, Vol. 38, No. 11, pp.39-41.

[Mizoguchi, 00] Mizoguchi R., Kozaki K., Sano T. et Kitamura Y. (2000) 'Construction and Deployment of a Plant Ontology', International Conference on Knowledge Engineering and Knowledge Management, pp.113-128, Springer Berlin Heidelberg.

[Moens, 06] Moens, M. F. (2006) 'Automatic indexing and abstracting of document texts', Vol. 6, livre, Springer Science & Business Media.

[Mohammed, 14] Mohammed, O. B. (2014) 'Application of Multimedia Technology in Database and IT Service Management', International Journal of Computer Applications, Vol. 103, No. 7.

[Moreira, 04] Moreira, A., Alvarenga, L. et de Paiva Oliveira, A. (2004) 'Thesaurus and ontology: A study of the definitions found in the computer and information science literature, by means of an analytical-synthetic method', Knowledge organization, Vol. 31, No. 4, pp.231-244.

[Mortensen, 13] Mortensen, J.M. (2013) 'Crowdsourcing ontology verification', In International Semantic Web Conference, pp. 448-455, Springer Berlin Heidelberg.

[Moulin, 12] Moulin, C., Largeron, C., Barat, C., Géry, M. et Ducottet, C. (2012) 'Apprentissage par analyse linéaire discriminante des paramètres de fusion pour la recherche d'information multimédia texte-image', Extraction et gestion des connaissances, EGC'2012, Hermann-Editions, pp. 357-368.

[Musen, 15] Musen, M.A. (2015) 'The protégé project: a look back and a look forward', International Journal of AI matters, Vol. 1, No. 4, pp.4–12.

[Mylonas, 08] Mylonas, P., Athanasiadis, T., Wallace, M., Avrithis, Y. et Kollias, S. (2008) 'Semantic representation of multimedia content: Knowledge representation and semantic indexing', Multimedia Tools and Applications, Vol. 39, No. 3, pp. 293-327.

[Neches, 91] Neches, R., Fikes, R. E., Finin, T., Gruber, T., Patil, R., Senator, T. et Swartout, W. R. (1991) 'Enabling technology for knowledge sharing', AI magazine, Vol. 12, No. 3, pp.36.

[Nefzi, 15] Nefzi, H., Farah, M. et Farah, I. R. (2015) 'Evaluation of the Taxonomic Consistency of Ontologies based on WordNet Hierarchical and Lexical Relations', In The International Congress for global Science and Technology, pp. 41.

[Niwattanakul, 13] Niwattanakul, S., Singthongchai, J., Naenudorn, E. et Wanapu, S. (2013) 'Using of Jaccard coefficient for keywords similarity', in Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2013, Vol. 1, pp.13–15.

[Noy, 04] Noy, N. F. et Musen, M. A. (2004) 'Ontology versioning in an ontology management framework', IEEE Intelligent Systems, Vol. 19, No. 4, pp.6-13.

[Noy, 01] Noy, N. F. et McGuinness, D. L. (2001) 'Ontology development 101: A guide to creating your first ontology', pp. 1-26.

[Noy, 00a] Noy, N. F., Fergerson, R. W. et Musen, M. A. (2000) 'The knowledge model of Protege-2000: Combining interoperability and flexibility', In International Conference on Knowledge Engineering and Knowledge Management, pp.17-32, Springer Berlin Heidelberg.

[Noy, 00b] Noy, N. F. et Musen, M. A. (2000) 'PROMPT: Algorithm and tool for automated ontology merging and alignment', in AAAI/IAAI, pp. 450–455

[Nwosu, 12] Nwosu, K. C., Thuraisingham, B. et Berra, P. B. (2012) 'Multimedia Database Systems: design and implementation strategies', Kluwer Academic Publishers, 377 pages.

[Ognyanov, 02] Ognyanov, D. et Kiryakov, A. (2002) 'Tracking changes in RDF (S) repositories', In International Conference on Knowledge Engineering and Knowledge Management, pp. 373-378, Springer Berlin Heidelberg.

[Parent, 96] Parent, C. et Spaccapietra, S. (1996) 'Integration de bases de données: Panorama des problèmes et des approches', Ingéniérie des Systèmes d'information, Vol. 4, No. LBD-ARTICLE-1996-002, pp.333-359.

[Parlangeau, 03] Parlangeau-Vallès, N., Farinas J., Fohr D., Illina I, Magrin Chagnolleau I, et al. (2003), 'Audio Indexing on the Web: A Preliminary Study of Some Audio Descriptors', 7th World Multiconference on Systematics, Cybernetics and Informatics.

[Patel-Schneider, 91] Patel-Schneider, P. F., McGuinness, D. L., Brachman, R. J. et Resnick, L.A. (1991) 'The CLASSIC knowledge representation system: Guiding principles and implementation rationale', ACM SIGART Bulletin, Vol. 2, No. 3, pp.108-113.

[Peeters, 00] Peeters, G., McAdams, S. et Herrera, P. (2000) 'Instrument sound description in the context of MPEG-7', In ICMC: International Computer Music Conference, pp. 166-169.

[Pellegrino, 04] Pellegrino F., Farinas J. et Rouas J-L. (2004) 'Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech', International Conference on Speech Prosody 2004, ISCA Special Interest Group on Speech Prosody (SproSIG), ISBN 2-9518233-1-2, pp. 517-520.

[Pérez, 06] Pérez, J., Arenas, M. et Gutierrez, C. (2006) 'Semantics and Complexity of SPARQL', International semantic web conference, pp. 30-43Springer Berlin Heidelberg.

[Petridis, 06] Petridis K., Anastasopoulos D., Saathoff C., Timmermann N., Kompatsiaris Y. et Staab, S. (2006) 'M-ontomat-annotizer: image annotation linking ontologies and multimedia low-level features', In: Gabrys B, Howlett RJ, Jain LC (eds) KES (3), Lecture notes in computer science, vol 4253. Springer, pp 633–640

[Picard, 95] Picard, R. W. et Minka, T. P. (1995) 'Vision texture for annotation', Multimedia systems, Vol. 3, No. 1, pp.3-14.

[Pierra, 05] Pierra, G., Hondjack, D., Ameur, Y. A. et Bellatreche, L. (2005) 'Bases de données à base ontologique. Principe et mise en œuvre', Ingénierie des systemes d'information, Vol. 10, No. 2, pp. 91-115.

[Pinquier, 04] Pinquier, J. (2004) 'Indexation sonore: recherche de composantes primaires pour une structuration audiovisuelle', Thèse de doctorat, Université Paul Sabatier-Toulouse III.

[Pottinger, 00] Pottinger, R. et Levy, A. Y. (2000) 'A Scalable Algorithm for Answering Queries Using Views', In VLDB, pp. 484-495.

[Poveda-Villalón, 12] Poveda-Villalón, M., Suárez-Figueroa, M. C. et Gómez-Pérez, A. (2012) 'Validating ontologies with oops!', In International Conference on Knowledge Engineering and Knowledge Management, pp. 267-281, Springer Berlin Heidelberg.

[Quilitz, 08] Quilitz, B. et Leser, U. (2008) 'Querying distributed RDF data sources with SPARQL', European Semantic Web Conference, Springer Berlin Heidelberg, pp. 524-538.

[Rahm, 11] Rahm, E. (2011) 'Towards large-scale schema and ontology matching', In Schema matching and mapping, pp. 3-27, Springer Berlin Heidelberg.

[Rajpathak, 13] Rajpathak, D. G. (2013) 'An ontology based text mining system for knowledge discovery from the diagnosis data in the automotive domain', Computers in Industry, Vol. 64, No. 5, pp.565-580.

[Ramasubramanian, 13] Ramasubramanian, C. et Ramya, R. (2013) 'Effective pre-processing activities in text mining using improved porter's stemming algorithm', International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No. 12, pp. 4536- 4538.

[Rani, 13] Rani, K. et Sharma, R. (2013) 'Study of different image fusion algorithm', International journal of Emerging Technology and advanced engineering, Vol. 3, No. 5, pp.288-291.

[Robertson, 76] Robertson, S. E. et Sparck Jones, K. (1976) 'Relevance weighting of search terms', Journal of the American Society for Information Sciences, Vol.27, No. 3, pp. 129-146.

[Rousset, 02] Rousset, M.C., Bidault, A., Froidevaux, C., Gagliardi, H., Goasdoué, F., Reynaud, C. et Safar, B. (2002) 'Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes: le projet PICSEL', Journal of Information Interaction Intelligence, Vol. 2, No. 1, pp.1–50.

[Rougui, 07] Rougui, J. E., Gelgon, M., Aboutajdine, D., Mouaddib, N. et Rziza, M. (2007) 'Organizing Gaussian mixture models into a tree for scaling up speaker retrieval', Pattern recognition letters, Vol. 28, No. 11, pp. 1314-1319.

[Sabri, 13] Sabri, A. M., Boonaert, J., Lecoeuche, S. et Mouaddib, E. (2013) 'Caractérisation spatio-temporelle des co-occurrences par ACP à noyau pour la classification des

actions humaines', In GRETSI'13.

[Safar, 09] Safar, B. et Reynaud, C. (2009) 'Alignement d'ontologies basé sur des ressources complémentaires Illustration sur le système TaxoMap', Technique et Science Informatiques, Vol. 28, No. 10, pp.1211-1232.

[Sagot, 08] Sagot, B. et Fišer, D. (2008) 'Construction d'un wordnet libre du français à partir de ressources multilingues', In TALN 2008-Traitement Automatique des Langues Naturelles.

[Salton, 89] Salton, G. (1989) 'Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer', Addison-Wesley, 530 pages.

[Salton, 88] Salton, G. et Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', Information processing & management, Vol. 24, No. 5, pp.513-523.

[Salton, 71] Salton, G. (1971) 'A comparison between manual and automatic indexing methods', Journal of American Documentation, Vol. 20, No. 1, pp.61–71.

[Sebe, 07] Sebe, N. et Tian, Q. (2007), 'Personalized multimedia retrieval: the new trend?', In Proceedings of the international workshop on Workshop on multimedia information retrieval, pp. 299-306, ACM.

[Shvaiko, 13] Shvaiko, P. et Euzenat, J. (2013) 'Ontology matching: state of the art and future challenges', IEEE Transactions on knowledge and data engineering, Vol. 25, No. 1, pp.158-176.

[Simperl, 09] Simperl, E. (2009) 'Reusing ontologies on the Semantic Web: A feasibility study', Data & Knowledge Engineering, Vol. 68, No. 10, pp.905-925.

[Singh, 14] Singh, M. et Kumar, K. (2014) 'Concept based automatic ontology generation from domain specific text', In International Conference of Soft Computing Techniques for Engineering and Technology, ICSCTET, pp. 1-5, IEEE.

[Smeaton, 12] Smeaton, A. (2012) 'Information retrieval and hypertext', Livre, Springer Science & Business Media, 273 pages.

[Stolze, 03] Stolze, K. (2003) 'SQL/MM Spatial-The Standard to Manage Spatial Data in a Relational Database System', In BTW, Vol. 2003, pp. 247-264.

[Smeulders, 00] Smeulders, A. W., Worring, M., Santini, S., Gupta, A. et Jain, R. (2000) 'Content-based image retrieval at the end of the early years', IEEE Transactions on pattern analysis and machine intelligence, Vol. 22, No.12, pp.1349-1380.

[Sellami, 14] Sellami, M., Gammoudi, M. M. et Hacid, M. S. (2014) 'Secure data integration: a formal concept analysis based approach', In International Conference on Database and Expert Systems Applications, pp. 326-333, Springer International Publishing.

[Shah, 14] Shah, M. M., et Patwal, P. S. (2014) 'Multi-dimensional image indexing with R-tree', International Journal of Innovations & Advancement in Computer Science, Vol. 3, No. 1, pp. 38-42.

[Sharma, 15] Sharma, D., Arora, U., Suri, P. et Tripathi, R. (2015) 'Better Scheme For Multimedia Compression Using Random Discrete Fractional Fourier Transform For Jpeg 2000 Standard', i-manager's Journal on Image Processing, Vol. 2, No. 1, pp.28.

[Sheth, 92] Sheth, A. P. et Kashyap, V. (1992) 'So far (schematically) yet so near (semantically) ', In Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), pp. 283-312.

[Sheth, 90] Sheth, A. P. et Larson, J. A. (1990) 'Federated database systems for managing distributed, heterogeneous, and autonomous databases', ACM Computing Surveys, CSUR, Vo. 22, No. 3, pp.183-236.

[Sowa, 95] Sowa J. (1995) 'Top-level ontological categories', International Journal of Human and Computer Studies, Vol. 43, No. 5-6, pp. 669-685.

[Sowa, 84] Sowa, J.F. (1984) 'Conceptual Structures: Information Processing in Mind and

Machine', Addison-Wesley, ISBN 0-201-14472-7.

[Spielmann, 10] Spielmann, Y. (2010) 'Video: the reflexive medium', The MIT Press, 384 pages.

[Stojanovic, 04] Stojanovic, L. (2004) 'Methods and tools for ontology evolution', Thèse de doctorat, Université de Karlsruhe, Allemagne, 249 pages.

[Strapparava, 04] Strapparava, C. et Valitutti, A. (2004) 'WordNet Affect: an Affective Extension of WordNet', In LREC, Vol. 4, pp. 1083-1086.

[Studer, 98] Studer, R., Benjamins, V. R. et Fensel, D. (1998) 'Knowledge engineering: principles and methods', Data & knowledge engineering, Vol. 25, No. 1-2, pp.161-197.

[Stoilos, 05] Stoilos, G., Stamou, G. et Kollias, S. (2005) 'A string metric for ontology alignment', In International Semantic Web Conference, pp. 624-637, Springer Berlin Heidelberg.

[Su, 12] Su, Z., Gao, J., Wang, Z., Dang, C., Shi, L. et Hu, G. (2012) 'Large object data storage technology analysis and integrated application', 2nd International Conference on Science and Network Technology, ICCSNT, pp. 1807-1812, IEEE.

[Subrahmanian, 95] Subrahmanian, V. D., Adali, S., Brink, A., Emery, R., Lu, J. J., Rajput, A. et. (1995) 'HERMES: A heterogeneous reasoning and mediator system', http://www.cs.umd.edu/projects/hermes/.

[Sultan, 13] Sultan, T.I., Nasr, M.M., Khedr, A.E. et Ismail, W.S. (2013) 'Semantic conflicts reconciliation (SCR): a framework for detecting and reconciling data-level semantic conflicts', International journal research and application, Vol. 3, No. 1, pp.766–773.

[Sun, 14] Sun, B., Luo, W. S., Du, L. B. et Lu, Q. (2014) 'Storage Model Based on Oracle InterMedia for Surveillance Video', Applied Mechanics and Materials, Vol. 644, pp. 3318-3321.

[Sure, 03] Sure, Y., Angele, J. et Staab, S. (2003) 'OntoEdit: Multifaceted inferencing for ontology engineering', In Journal on Data Semantics I, pp. 128-152, Springer Berlin Heidelberg.

[Sure, 02] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R. et Wenke, D. (2002) 'OntoEdit: Collaborative ontology development for the semantic web', In International Semantic Web Conference, pp. 221-235, Springer Berlin Heidelberg.

[Suwanmanee, 05] Suwanmanee, S., Benslimane, D., Champin, P.A. and Thiran, P. (2005) 'Wrapping and integrating heterogeneous databases with OWL', in Proceeding of the 7th International Conference on Enterprise Information Systems, USA, pp.11–18.

[Swain, 91] Swain, M. J. et Ballard, D. H. (1991) 'Color indexing', International journal of computer vision, Vol. 7, No. 1, pp.s11-32.

[Tamine, 00] Tamine, L. (2000) 'Optimisation de requêtes dans un système de recherche d'information approche basée sur l'exploitation de techniques avancées de l'algorithmique génétique', Thèse de doctorat, Université Paul Sabatier-Toulouse III.

[Tarakanov, 15] Tarakanov, O. V. (2015) 'A Comparative Research into Ways to Store Binary Large Objects in the ORACLE Databases', Trudy SPIIRAN, Vol. 42, pp.77-89.

[Teitsma, 14] Teitsma, M., Sandberg, J., Schreiber, G., Wielinga, B. et van Hage, W. R. (2014) 'Engineering ontologies for question answering', Applied Ontology, Vol. 9, No.1, pp.1-25.

[Torresani, 10] Torresani, L., Szummer, M. et Fitzgibbon, A. (2010) 'Efficient object category recognition using classemes', In European conference on computer vision, Springer Berlin Heidelberg, pp. 776-789.

[Ullman, 97] Ullman, J. D. (1997) 'Information integration using logical views', In International Conference on Database Theory, pp. 19-40, Springer Berlin Heidelberg.

[Uschold, 96] Uschold, M. et Gruninger, M. (1996) 'Ontologies: principles, methods and applications', Journal of Knowledge Engineering Review, Vol. 11, No. 2, pp.93–136.

[Uschold, 95] Uschold, M. et King, M. (1995) 'Towards a Methodology for Building Ontologies', In Workshop on Basic Ontological Issues in Knowledge Sharing, Held in Conjunction with IJCAI-95, pp.1-13.

[Uzdanaviciute, 11] Uzdanaviciute, V. et Butleris, R. (2011) 'Ontology-based Foundations for Data Integration', In BUSTECH 2011, The First International Conference on Business Intelligence and Technology, pp. 34-39.

[Van-Heijst, 97] Van-Heijst G., Schreiber A. et Wielinga B. J. (1997) 'Using Explicit Ontologies in KBS Development', International Journal of Human and Computer Studies, Vol. 46, No. 2-3, pp.183-292.

[Varelas, 05] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G.M. et Milios, E.E. (2005) 'Semantic similarity methods in WordNet and their application to information retrieval on the web', in Proceedings of the 7th annual ACM international Workshop on Web Information and Data Management, WIDM 2005, pp.10–16.

[Vassiliadis, 03] Vassiliadis, P., Simitsis, A., Georgantas, P. et Terrovitis, M. (2003) 'A Framework for the Design of ETL Scenarios', International Conference on Advanced Information Systems Engineering, pp. 520-535, Springer Berlin Heidelberg.

[Venturini, 14] Venturini, R. (2014) 'Experiments on Compressed Full-Text Indexing', Compressed Data Structures for Strings, Vol. 4, Atlantis Press, pp. 61-88.

[Vidal, 13] Vidal, V. M., De Macêdo, J. A., Pinheiro, J. C., Casanova, M. A. et Porto, F. (2013) 'Query processing in a mediator based framework for linked data integration', In Web-Based Multimedia Advancements in Data Communications and Networking Technologies, pp. 98-116, IGI Global.

[Visser, 04] Visser, U. (2004) 'General approach of Buster', in Visser, U. (Eds.): Intelligent Information Integration for the Semantic Web, pp.37–51, Springer Berlin Heidelberg.

[Visser, 99] Visser, P. R., Jones, D. M., Beer, M. D., Bench-Capon, T. J. M., Diaz, B. M., et Shave, M. J. R. (1999) 'Resolving ontological heterogeneity in the KRAFT project', In International Conference on Database and Expert Systems Applications, pp. 668-677, Springer Berlin Heidelberg.

[Wache, 01] Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. et Hübner, S. (2001) 'Ontology-based integration of information-a survey of existing approaches' in IJCAI 2001: Workshop: Ontologies and Information Sharing, Seattle, USA, pp.108–117.

[Wahlster, 86] Wahlster, W. et Kobsa, A. (1986) 'Dialogue-based user models', Proceedings of the IEEE, Vol. 74, No. 7, pp. 948-960.

[Wang, 06] Wang, J. et Chang, C. I. (2006) 'Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis', IEEE transactions on geoscience and remote sensing, Vol. 44, No. 6, pp.1586-1600.

[Weibel, 97] Weibel, S. (1997) 'The Dublin Core: A Simple Content Description Model for Electronic Resources', Bulltin of the Association for Information Science and Technology, Vol.24, No.1, pp. 9–11.

[Welty, 01] Welty C. et Guarino, N., (2001) 'Supporting Ontological Analysis of Taxonomic Relationships', Data et Knowledge Engineering, Vol. 39, No. 1, pp. 51-74.

[Wiederhold, 92] Wiederhold G. (1992) 'Mediators in the architecture of future information systems', IEEE computers, Vol. 25, No. 3, pp. 38-49.

[Woelk, 87] Woelk, D. et Kim, W. (1987) 'An extensible framework for multimedia information management', Database Engineering, Vol. 10, No. 2, pp. 55-61.

[Wu et Palmer, 94] Wu, Z. et Palmer, M. (1994) 'Verbs semantics and lexical selection', in ACL 1994:

Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, pp.133–138.

[Xu, 04] Xu, L. et Embley, D. W. (2004) 'Combining the Best of Global-as-View and Local-as-View for Data Integration', In ISTA, Vol. 48, pp. 123-136.

[Yen, 93] Yen, M. M. et Scamell, R. W. (1993) 'A human factors experimental comparison of SQL and QBE', IEEE Transactions on Software Engineering, Vol. 19, No. 4, pp.390-409.

[Zghal, 07] Zghal, S., Yahia, S. B., Nguifo, E. M. et Slimani, Y. (2007) 'SODA: Une approche structurelle pour l'alignement d'ontologies OWL-DL', In Proceedings of the first French Conference on Ontology, JFO2007, Sousse, Tunisia.

[Zhang, 08] Zhang, Y., Wu, J. et Zhuang, Y. (2008) 'Personalized Multimedia Retrieval in CADAL Digital Library', Advances in Multimedia Information Processing-PCM, pp. 703-712.

[Zhang, 04] Zhang, D. et Lu, G. (2004) 'Review of shape representation and description techniques', Pattern recognition, Vol. 37, No. 1, pp.1-19.

[Zhang, 03] Zhang, D. et Lu, G. (2003) 'Evaluation of MPEG-7 shape descriptors against other shape descriptors', Multimedia Systems, Vol. 9, No. 1, pp.15-30.

[Zhou, 10] Zhou, X., Yu, K., Zhang, T. et Huang, T. S. (2010) 'Image classification using super-vector coding of local image descriptors', In European conference on computer vision, Springer Berlin Heidelberg, pp. 141-154.

[Zloof, 77] Zloof, M. M. (1977) 'Query-by-example: A data base language', IBM systems Journal, Vo.16, No. 4, pp.324-343.