AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n°92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992
http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

Université de Sfax
Ecole Doctorale des Sciences Economiques, Gestion et Informatique
Faculté des Sciences Economiques et de Gestion de Sfax

Université Toulouse I Capitole
Ecole Doctorale Mathématique, Informatique et Télécommunication de Toulouse

THESE DE DOCTORAT

En vue de l'obtention du

DOCTORAT EN INFORMATIQUE

Présentée et soutenue par :

Salma BEN MEFTAH

Le 05/09/2017

STRUCTURATION SEMANTIQUE DE DOCUMENTS XML CENTRES-DOCUMENTS

Jury :

Mme. Hanêne Ben-Abdallah	Professeur d'informatique - FSEG - Sfax	Président
M. Jamel FEKI	Professeur d'informatique - FSEG - Sfax	Directeur de thèse
Mme. Chantal Soulé-Dupuy	Professeur d'informatique - IRIT - Toulouse	Directeur de thèse
Mme. Elisabeth MURISASCO	Professeur d'informatique - Université du Sud Toulon- Toulon	Rapporteur
M. Yahia SLIMANI	Professeur d'informatique, Institut Supérieur des Arts Multimédia de la Manouba - Tunis	Rapporteur
M. Mounir BEN AYED	Professeur d'informatique - faculté des Sciences Sfax	Membre

Résumé

La numérisation des documents et le développement des technologies Internet ont engendré une augmentation permanente du nombre de documents et de types de documents disponibles. Face à cette masse documentaire, XML (eXtensible Markup Language) s'est imposé comme format standard de structuration et d'échange de documents. Ainsi, un nombre de plus en plus important de documents devient disponible sous ce format. Ces documents XML peuvent être classés en deux types : les documents XML orienté-données et les documents XML orienté-textes.

Les documents XML orienté-données sont constitués d'un ensemble d'éléments généralement courts et précis et sont similaires aux données relationnelles. Nous constatons que les balises utilisées pour ce type de documents décrivent généralement d'une manière précise le contenu, et offrent la sémantique basique nécessaire à la description de l'information (Exemples de balises : Article, Client, Quantité, Prix).

A contrario, les documents XML orienté-textes sont riches en texte et utilisent des balises qui reflètent la plupart du temps un découpage (structurel) logique (exemples de balises : Contenu, Section, Paragraphe). Malheureusement, ces balises n'ont qu'une très pauvre vocation sémantique.

Partant de cette constatation, le développement d'approches supportées par des outils automatisés permettant de décrire la sémantique des documents XML orientés-textes devient un besoin urgent, voire une nécessité pour certains usages. Dans ce contexte, nous proposons *une approche de structuration sémantique des documents XML à partir de leurs structures logiques et de leurs contenus*. Elle construit une arborescence de concepts.

Cette approche de structuration sémantique passe par quatre phases : 1) Extraction des termes des contenus des documents en utilisant des techniques de recherche d'information ; 2) Détermination d'une taxonomie[1] qui sera affectée au document, c'est-à-dire celle qui correspond au mieux à sa sémantique (cette étape se base sur une démarche de pondération d'un ensemble de taxonomies candidates) ; 3) Affectation, à chaque élément feuille de la structure logique du document, du concept le plus significatif à partir de la taxonomie retenue ; 4) Inférence de concepts aux éléments non feuilles du document.

Notre approche de structuration sémantique des documents se base sur l'indexation sémantique et diffère des autres travaux par : 1) Le choix d'une taxonomie appropriée pour

[1] C'est une ressource sémantique dont les concepts sont reliés entre eux par des liens hiérarchiques.

chaque document, il s'agit de déterminer la taxonomie qui décrit au mieux la sémantique du document, et 2) La pondération des concepts extraits de manière à donner plus d'importance aux concepts les plus spécifiques car nous partons du constat suivant : plus le niveau auquel se situe le concept est bas dans la hiérarchie, plus l'information qu'il apporte est fine et ciblée.

Pour exploiter ces structures sémantiques, nous avons *étendu le méta-modèle d'entrepôts de documents* pour assurer leur stockage. De plus, nous avons introduit le *concept de méta-document* afin de permettre l'interrogation de ces structures sémantiques. Enfin, pour évaluer nos propositions, nous avons mené un ensemble d'expérimentations sur la collection de documents *XML ImageCLEFMed* 2010 en utilisant la ressource sémantique *MeSH (*NML's Medical Subject Headings). Les résultats obtenus montrent que l'algorithme de pondération des concepts des taxonomies qui a été proposé permet de sélectionner avec précision la taxonomie pertinente pour un document donné et, en conséquence, les concepts pertinents à affecter aux éléments feuilles de la structure sémantique de ce document.

Remerciements

A la mémoire de mon père

Je dédie ce rapport essentiellement à la mémoire de mon père, que Dieu le prenne dans son vaste paradis, car il a toujours souhaité me voir terminer mes études.

A ma mère

Un grand merci plein d'affection à ma mère qui m'a accompagnée durant toute ma vie en m'accueillant sous son toit et qui est toujours là pour moi.

A mon mari

Je vous suis reconnaissante pour l'affection que vous me témoignez. Que Dieu me donne le pouvoir de ne pas oublier mon devoir de mémoire pour toute l'aide et l'attachement dont vous avez fait preuve à mon égard.

A mon frère et ma sœur

Je remercie aussi ma sœur et mon frère qui m'ont encouragée par leurs soutiens moral et leurs affections et qui n'ont jamais cessé de me soutenir durant la période de mes études.

A mes directeurs de thèses

J'ai l'agréable plaisir de remercier vivement mes deux directeurs de thèse Monsieur Jamel Feki, professeur à la Faculté des Sciences Économiques et de Gestion de l'Université de Sfax, et Madame Chantal Soulé-Dupuy, professeur à l'Université Toulouse 1 Capitole pour leur soutien et collaboration de tous les instants. Leur aide et leur disponibilité continues, ainsi que leurs précieuses remarques constructives qui ont grandement contribué à bien mener ce travail et améliorer la qualité de ce mémoire de thèse.

Un merci tout particulier à Monsieur Kaïs Khrouf, maître assistant à l'École Nationale d'Électronique et des Télécommunications de Sfax pour son aide continue, ses conseils, ses encouragements et pour le temps précieux qu'il m'a consacré pour suivre en détail toutes les étapes de réalisation de ce travail.

Aux membres de Jury

Je remercie Monsieur Slimani Yahia, professeur de l'Université de la Manouba, Institut Supérieur des Arts Multimédia de la Manouba, Laboratoire LISI, pour avoir accepté de participer à mon jury en tant que rapporteur.

Je voudrais aussi remercier Monsieur Mounir Ben Ayed, professeur de la faculté des Sciences Sfax, pour avoir accepté d'être membre de mon jury en tant que membre.

Je remercie également Madame Elisabeth Murisasco, professeur d'informatique - Université du Sud Toulon-Toulon, pour avoir accepté d'être rapporteur de mon jury.

A mes amis

Pour tous mes amis et tous ceux que je n'ai pas pu citer, veuillez trouver dans ce Travail l'expression de grande considération. Merci de votre précieuse amitié.

Salma Ben Meftah

Sommaire

Listes des figures

Listes des tableaux

Introduction générale

Sommaire

1 Contexte et problématique

Le développement d'Internet a largement contribué à augmenter le nombre de documents et les volumes de données disponibles et échangés sous forme numérique via les réseaux, le Web en particulier. La conséquence de cette augmentation est que l'utilisateur éprouve de plus en plus de difficultés pour retrouver l'information pertinente dont il a besoin dans cette masse de documents.

En règle générale, pour qu'un document soit retourné à un utilisateur en réponse à une requête, ce document doit contenir au moins un terme parmi l'ensemble des termes décrivant le besoin. Cependant, pour l'utilisateur, un document pourrait être pertinent même s'il ne contient pas les termes de la requête. Par exemple, si l'utilisateur souhaite chercher des documents contenant le terme « système d'exploitation » alors tout document contenant *Windows*, *Unix* ou *Vista* devrait être considéré comme pertinent et ajouté au résultat de la recherche exacte par le terme « *système d'exploitation »*.

D'un autre côté, faciliter l'échange de documents à travers le Web, *XML* est devenu le format standard communément utilisé pour décrire ces documents. Pour accéder à ces documents XML, le processus classique de recherche d'information (RI) repose sur l'analyse de leur contenu textuel alors que la structure logique n'intervient que rarement. De plus, cette structure logique est complètement dépourvue de sémantique. Or, la restitution d'un résultat pertinent d'interrogation mériterait d'exploiter la sémantique, actuellement absente, des documents. D'où la nécessité d'approches permettant de décrire la structure sémantique des documents XML.

Dans ce contexte, les problématiques abordées dans la cadre de cette thèse, sont : 1) La détermination d'une structure sémantique des documents XML, et 2) L'exploitation de ces structures dans un processus d'interrogation.

2 Contributions

Les travaux présentés dans ce mémoire de thèse se situent dans le contexte de l'indexation sémantique de documents XML centrés-textes pour la génération automatique de leur structure sémantique.

Dans la littérature traitant de sémantique de documents, nous pouvons recenser plusieurs définitions pour la notion de « structure sémantique ». Abascal (Abascal R., 2005) définit la structure sémantique comme étant « *un ensemble de balises sémantiques représentant des concepts associés entre eux par des relations* ».

Dans notre contexte, nous considérons que « *La structure sémantique d'un document XML est un arbre dérivée de sa structure logique et dont les nœuds sont des concepts qui*

caractérisent le contenu textuel du document et extraits à partir d'une ressource sémantique associée au document ».

Dans ce sens, nos contributions portent sur les quatre volets suivants :

1) La pondération automatique des concepts des taxonomies de manière à accorder plus d'importance aux concepts les plus spécifiques (situés en bas de la hiérarchie,) car nous jugeons qu'un concept-fils représente une information plus précise que son concept-père.

2) La proposition d'une approche pour la détermination de la structure sémantique d'un document *XML* en se référant à une ressource sémantique. L'objectif principal est de capter la sémantique d'un document et de représenter ses différentes parties par le biais d'un ensemble de concepts organisés sous forme d'un arbre dérivé de sa structure logique. Il s'agit, tout d'abord, de déterminer la ressource sémantique la plus appropriée pour un document, parmi un ensemble de taxonomies. Ensuite, les concepts de la taxonomie retenus seront affectés aux éléments feuilles et propagés aux éléments non feuilles de l'arbre sémantique.

3) L'exploitation et l'utilisation des structures sémantiques déduites lors de la contribution précédente. Tout d'abord, nous proposons l'intégration de la structure sémantique dans l'entrepôt de documents et ce par extension du méta-modèle de l'entrepôt (Khrouf K., 2004). Ensuite, nous définissons un méta-document XML qui a pour but de décrire la sémantique d'un document XML ; ce méta-document sera utilisé ultérieurement pour interroger les structures (logiques et sémantiques) et les contenus des documents XML (Ben Meftah S. et al., 2015).

4) L'expérimentation pour valider notre approche et montrer notamment l'apport de la pondération automatique des taxonomies. Ainsi, nous avons utilisé : 1) Le thésaurus *MeSH* spécialisé dans le domaine médical comme ressource sémantique, et 2) La collection de documents XML*ImageCLEFMed 2010* comme base de test.

3 Organisation de la thèse

Ce mémoire de thèse est organisé en cinq chapitres :

- Le **Chapitre 1** présente les principaux concepts de la norme XML et des entrepôts de documents. Il introduit également les principes de base de l'indexation classique (dite « sac de mots ») et celle de l'indexation sémantique.
- Le **Chapitre 2** dresse un état de l'art des travaux traitant de l'indexation sémantique. Nous classons les différents travaux de la littérature selon le type de ressource sémantique utilisée : généralisée ou spécialisée.

– Le **Chapitre 3** décrit notre approche pour la construction de la structure sémantique d'un document *XML*. Il s'agit d'une structure complémentaire à la structure logique pour décrire les concepts traités dans le document.

– Le **Chapitre 4** traite l'exploitation de la structure sémantique. Pour ce faire, nous avons étendu le méta-modèle d'entrepôts de documents pour supporter ces structures sémantiques et nous avons introduit le concept de méta-document pour permettre leur interrogation.

– Le **Chapitre 5** s'intéresse à l'expérimentation et à l'analyse des résultats obtenus. L'expérimentation est effectuée avec 1) Le thésaurus *MeSH*, spécialisé dans le domaine médical pour indexer et rechercher des articles, et sur 2) La collection de documents *ImageCLEFMed 2010* (Ben Meftah S. et al., 2014).

Chapitre 1

Concepts de base pour l'entreposage et l'indexation de documents

Sommaire

1.1 Introduction

L'essor des technologies de l'information, avec l'avènement d'Internet et des réseaux, a augmenté de manière considérable le volume d'informations disponibles et accessibles par le grand public. De façon similaire, les entreprises font face à un volume croissant de données, documents… manipulés par leur système d'information ; cette masse rend difficile l'exploitation des informations. Dans ce contexte, XML s'est imposé comme format standard de documents. Ces documents sont décrits par des structures différentes telles que la structure logique, la structure physique, la structure temporelle, etc. Dans les *SRIs* «*Systèmes de Recherche d'Information »* classiques, un document est considéré comme un ensemble de mots simples (appelés aussi *sac de mots* ou *termes d'indexation* (Salton G.et McGill M.J., 1983) (Baeza-yates R. et al., 1999)) qui sont utiles pour l'indexation classique. Dans ce type d'indexation, les mots sont considérés comme des graphies sans sémantique. Les seules informations utilisées concernant ces mots sont leurs fréquences d'apparition dans les documents (approches statistiques). Conséquemment, ces systèmes ne prennent pas en considération le sens du mot (Genest et al., 2005), de plus ils ne distinguent pas les mots selon leurs contextes d'apparition. Afin de remédier à ces limites, plusieurs travaux se sont intéressés à la prise en compte de l'aspect sémantique des termes d'indexation. Ce type d'indexation est appelé indexation sémantique ou conceptuelle.

Dans ce chapitre, nous présentons, dans une première section, la norme *XML* en tant que standard de représentation et d'échange de documents. Nous décrivons également les différentes structures et catégories de documents XML (notamment les documents orientés-textes et orientés-données). Ensuite, nous définissons l'entreposage de documents. Dans la section suivante, nous définissons la notion d'indexation de documents textuels ainsi que les différents types d'indexation. Enfin, nous nous focalisons sur l'indexation sémantique et nous détaillons les différentes ressources sémantiques utilisées pour ce type d'indexation.

1.2 La norme XML

XML "eXtensible Markup Language" (W3C, 1999) est un langage de description et d'échange de documents structurés. A l'aide d'un système de balisage, *XML* permet de marquer les éléments qui composent la structure d'un document et les relations entre ces éléments. Son objectif est de définir un formalisme permettant d'échanger facilement des documents complexes sur le Web.

Un document *XML* contient une ou plusieurs unités d'information où chaque unité d'information est une chaîne de caractères délimitée par deux balises (une balise est une suite de caractères encadrée par " < " et " > "). Chaque unité d'information peut être elle-même composée d'autres sous-unités.

Chaque document *XML* peut ainsi être considéré comme un ensemble d'éléments (unités d'information) organisés hiérarchiquement (selon une organisation logique) et/ou un enchaînement temporel d'éléments (organisation temporelle) et/ou un agencement d'un ensemble de blocs (organisation physique), etc. A chaque organisation correspond un type particulier de structure.

Dans cette section, nous nous intéressons au contenu du document XML, les différentes structures qui peuvent lui être associées ainsi que les différentes catégories de documents XML.

1.2.1 Composition et structuration de documents XML

Nous commençons cette section par décrire la composition d'un document XML :

- Un prologue contenant la première ligne du document *XML*. Il donne des informations de traitement. Voici un exemple de prologue :

```
<?xml version="1.0" encoding="UTF-8"?>
```

- Un ensemble de balises qui décrivent la *structure logique*. Selon (Fourel F., 1998), la structure logique reflète « *l'organisation explicite d'abstractions logiques représentant des parties d'un document* ». Cette structure logique permet un découpage de l'information d'un point de vue hiérarchique et logique selon un principe de décomposition plus ou moins fin. Ce mécanisme impose d'identifier de façon non ambigüe les granules d'information composant le document ;
- Le contenu proprement dit du document encadré par des balises ;
- Des commentaires et des instructions de traitement dont la présence est facultative.

Selon (Roisin C., 1999), la structure logique des documents s'appuie sur trois entités :

- Les éléments de base non décomposables qui constituent le contenu ;
- Les éléments composites, obtenus par composition d'éléments de base ou d'autres éléments composites ;
- Les attributs qui peuvent être associés aux éléments pour leur adjoindre des informations supplémentaires.

La Figure 1 montre la structure logique d'un document *XML* nommée "*Paper*". Cette structure est composée de 3 éléments composites : *Paper*, *Conference* et *Abstract*. L'élément *Paper* est composé de deux éléments de base *Title* et *Author* et de deux éléments composites *Conference* et *Abstract*. A son tour, l'élément *Conference* est composé de deux éléments de base *Name* et *Year*. Enfin, l'élément *Abstract* est composé de deux éléments de base qui sont les deux paragraphes *P1* et *P2*.

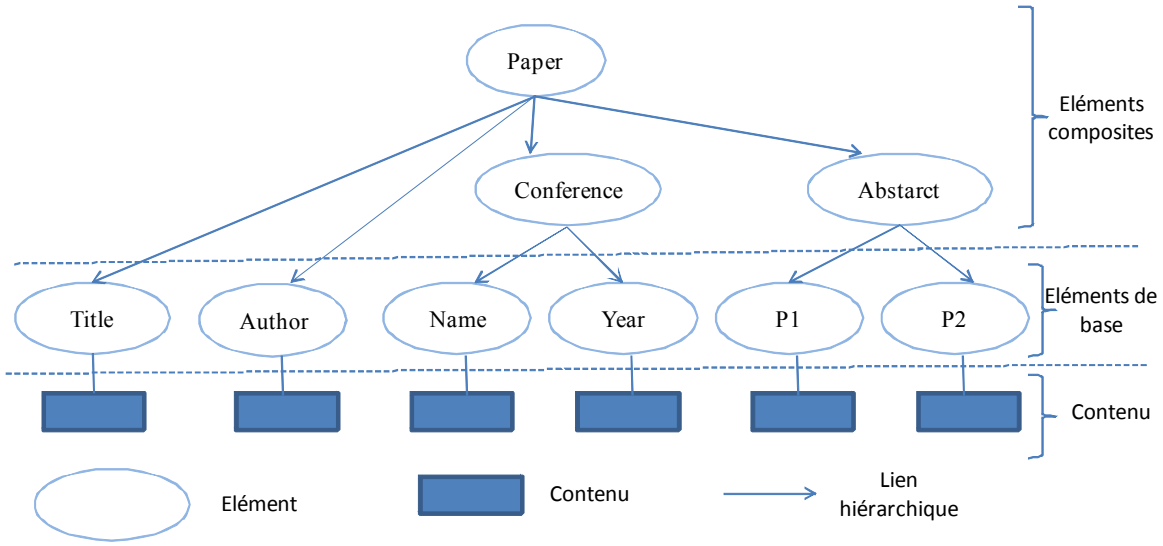


Figure 1. Structure logique du document "*Paper*".

En plus de sa structure logique, à un document XML peuvent être associés d'autres types de structures, à savoir :

- *La structure physique* : Le concept de structure physique est lié à la restitution du document sur un support physique (papier, écran, etc.), c'est-à-dire à l'organisation spatiale des éléments sur le support de restitution. Elle définit les différentes zones d'un document ainsi que leurs caractéristiques et décrit l'agencement de ces zones. Ainsi, selon la restitution souhaitée, un découpage du document en zones doit être effectué. Pour cela, un ensemble de règles de présentation est utilisé pour traduire cette structure telles qu'une succession de lignes, de paragraphes, de colonnes, de pages, de caractéristiques typographiques, etc. Ces règles sont spécifiées par pavés (ou blocs) d'information. Tout comme la structure logique, la structure physique est présentée sous forme d'une arborescence de blocs. Sur un support papier, le découpage s'effectue par exemple page par page, colonne par colonne ou paragraphe par paragraphe ;
- *La structure temporelle :* Les documents audio et audiovisuel sont construits autour d'un axe temporel. Cet axe met en jeu les planifications et les enchaînements des différents éléments du contenu. La structure temporelle permet de traduire et décrire ces planifications et ces enchaînements.

1.2.2 Catégories dedocuments XML

Les documents XML peuvent être classés en deux grandes catégories (Ravat F. et al., 2010) : orientés-données ("*data-centric*") et orientés-textes ("*text-centric*").

1.2.2.1 Les documents XML orientés-données

Ces documents représentent des données fortement structurées comme le contenu d'une base de données relationnelle. Ce type de documents est principalement utilisé par les applications d'échange d'informations. Les balises utilisées dans ce genre de documents

décrivent généralement d'une manière précise son contenu, et annoncent la sémantique nécessaire à la description de l'information en question (par exemple <NomProd>…</NomProd>). La Figure 2 illustre un exemple de document XML orienté-données représentant une transaction de vente.

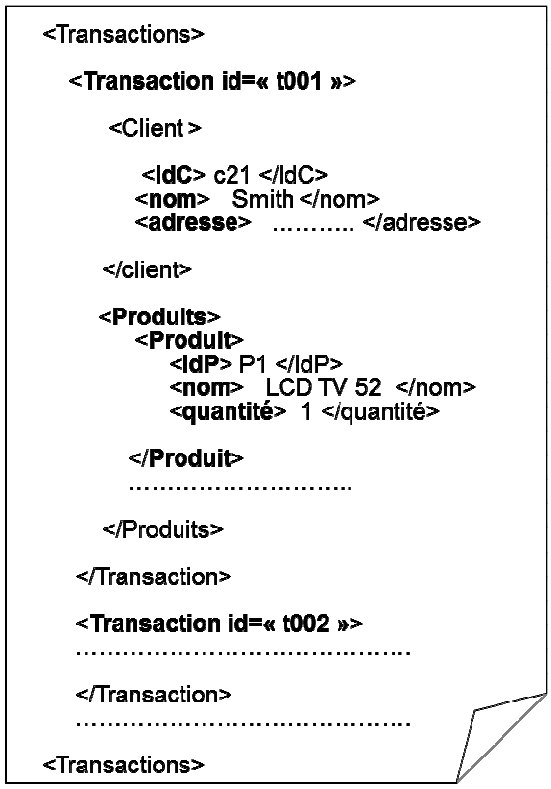


Figure 2. Exemple de document XML orienté-données pour une transaction de ventes de produits (Transactions.xml).

1.2.2.2 Documents XML orientés-textes

Les documents XML orientés-textes sont principalement composés de textes comme par exemple des versions électroniques de documents dits non-structurés ou semi-structurés selon le type des balises utilisées. Ces documents ont le plus souvent des structures hétérogènes. Les balises utilisées présentent généralement un découpage logique, d'où la notion de structure logique. Toutefois, ces balises de structuration logique n'ont pas une vocation sémantique.

On distingue donc deux types de balises dans les documents orientés-textes :

- Les balises de type « Attribut-Valeur », comme par exemple *NomProd*, *Quantité* (cf. Figure 2) ou encore *Name*, *Year* (cf. Figure 3).
- Les balises de type structurel, comme par exemple *Abstract* (cf. Figure 3).

La *Figure 3* présente un exemple de document XML orienté-textes décrivant les éléments d'un article de recherche.

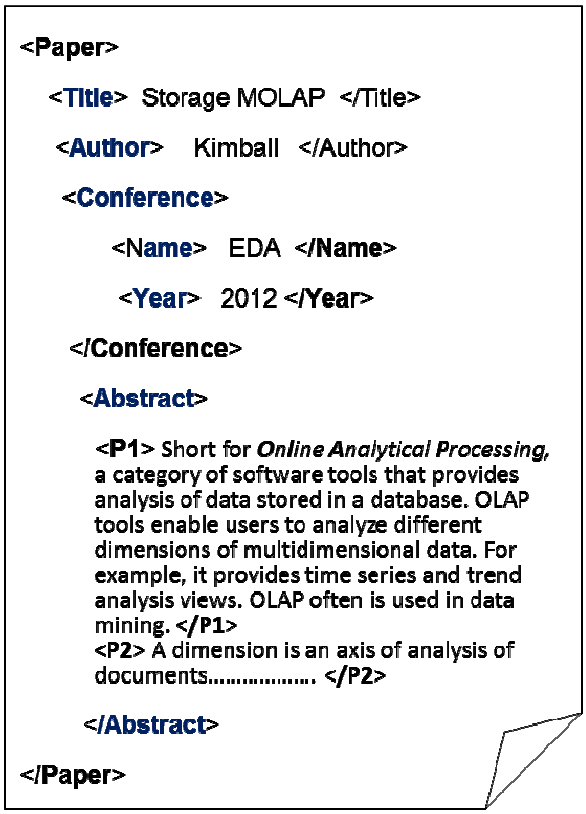


Figure 3. Exemple de document XML orienté-textes (paper.xml).

Notons que pour les documents *XML* orientés-données, l'ordre des éléments est généralement moins important que pour les documents orientés-textes. Par exemple, dans Transactions.xml (cf. Figure 2), mettre la quantité avant ou après le nom du produit dans une transaction de vente ne cause aucun problème, ce qui n'est pas le cas pour les paragraphes d'un document textuel. En effet, inverser l'ordre des paragraphes d'un document nuit à la compréhension de son contenu et peut même en modifier le sens.

Afin d'exploiter les contenus des documents (entre autres les documents XML), le concept d'entrepôt de documents a été proposé ; il vise à stocker et analyser les documents. Ce concept d'*entrepôt de documents* fera l'objet de la section suivante.

1.3 Les entrepôts de documents

Depuis de nombreuses années, les entreprises ont mis en place des systèmes leur permettant de mieux gérer les documents tels que les systèmes de Gestion Électronique des Documents(*GED*)[2], les systèmes de *Workflow*[3]ou encore les systèmes de *Groupware*[4].

[2]http://fr.wikipedia.org/wiki/Gestion_%C3%A9lectronique_des_documents

L'inconvénient majeur de ces systèmes réside dans leur caractère spécifique et ad hoc. En effet, chaque système manipule et conserve les documents dans un environnement spécifique qui lui est propre. Le concept d'entrepôt de documents est destiné à résoudre ce problème car il s'agit d'une approche plus générique et ouverte.

1.3.1 Définition de l'entrepôt de documents

Parmi les définitions proposées dans la littérature, nous avons retenu celles qui semblent être les plus courantes et les plus appropriées à notre travail.

Selon Sullivan (Sullivan D., 2001), « *L'entrepôt de documents doit être considéré comme un environnement fondé sur des standards permettant aux utilisateurs de capturer, d'analyser et de croiser un ensemble d'informations dans un contexte facilitant son accès et sa diffusion* ».

Dans le même sens, (Khrouf K., 2004) définit un entrepôt de documents comme étant « *Une source d'informations orientées-sujets, filtrées, intégrées, historiées et organisées comme support d'un processus de recherche, d'interrogation ou d'analyse* ».

Nous détaillons dans ce qui suit les termes de cette définition :

- *Orientées-sujet* : Les données d'un entrepôt s'organisent par sujet. Un sujet permet ainsi de rassembler les documents désignant un thème en se fondant sur leur structure et leur contenu, ainsi que sur leurs attributs. Cette différence de l'entrepôt permet de faciliter l'application des processus de recherche, d'interrogation et d'analyse des documents.
- *Filtrées* : Un entrepôt ne comprend que les documents sectionnés pertinents pour l'entreprise.
- *Intégrées* : Les données d'un entrepôt sont la conséquence de l'intégration de documents en provenance de diverses sources et de formats différents.
- *Historiées* : L'entrepôt doit permettre l'historisation des documents, c'est à dire préserver leurs différentes versions.
- *Recherche (ou interrogation) du contenu textuel* : Il s'agit de retrouver des documents ou tranches de documents en fonction de leurs contenus et en réponse à une requête exprimée en utilisant des mots-clés.
- *Interrogation du contenu factuel* : Elle consiste à utiliser un langage déclaratif pour consulter l'entrepôt et restituer des données factuelles et des métadonnées (des informations sur le contenu et les attributs des documents).

[3]http://fr.wikipedia.org/wiki/Workflow
[4] hp://fr.wikipedia.org/wiki/Groupware

- *Analyse* : Elle consiste à étudier les informations documentaires par leurs contenus et structures en utilisant les techniques *OLAP* (On-Line Analytical Processing) (Tournier R., 2007).

1.3.2 Objectifs d'un entrepôt de documents

L'entrepôt de documents est ainsi un catalogue central qui regroupe l'ensemble des documents sélectionnés pertinents pour un utilisateur ou un groupe d'utilisateurs, dans le cadre d'une activité ou pour aboutir à un objectif donné. Il intègre à la fois des données textuelles internes à l'organisation mais aussi des données issues de sources externes, comme le Web. La centralisation des documents dans l'entrepôt facilite leur accès et leur analyse. Elle permet également de découvrir de nouvelles informations ou connaissances à partir de documents non liés entre eux a priori.

Un entrepôt de documents doit permettre :

- *L'acquisition des documents* : Les documents sont assemblés à partir des systèmes de gestion des informations de l'entreprise (*GED*, Workflow, etc.) ou issus de sources externes et disséminées (Internet, bibliothèques numériques, etc.).
- *L'échange et la répartition de documents* dans un cadre d'interopérabilité de systèmes d'information par l'utilisation d'un format standard d'échange.
- *L'accès aux documents* : L'accès à l'ensemble des informations doit être simple et facile quels que soient les formats et les contenus. Ce type d'accès doit permettre de rechercher, retrouver et restituer des informations pertinentes selon différents modes possibles, telle que la recherche directe ou l'exploration, en tenant compte du contenu sémantique des documents.

Les nouvelles exigences dans le domaine des entrepôts de documents se tournent vers l'emploi de nouvelles pratiques et notamment vers l'exploitation de documents orientés-textes. Il faut pour cela : (i) aller au-delà du factuel et du numérique, et (ii) aller au-delà d'une simple exploitation textuelle de type « sac de mots ». Il faut alors pouvoir exploiter la sémantique contenue dans ces textes.

Pour ce faire et pour pouvoir exploiter la sémantique du contenu des documents, une tâche d'indexation est nécessaire ; elle détermine la représentation, par des descripteurs sémantiques, de chacun des documents. Chacune de ces représentations sera ensuite comparée à la représentation du besoin utilisateur afin de déterminer les documents susceptibles d'être pertinents. En fait, l'indexation est une étape capitale puisque la qualité de la restitution des informations résultera de la qualité de l'indexation. Elle est difficile à réaliser puisqu'il s'agit de définir les éléments considérés comme significatifs du contenu de chacun des documents. Il existe deux types d'indexation : l'*Indexation classique* de type « *Sac de mots* » et l'*Indexation sémantique* plus significative et efficace mais nécessitant plus d'efforts et de

ressources. Nous décrivons chacun de ces deux types d'indexation dans les deux sections qui suivent.

1.4 Indexation classique de type « *Sac de mots* »

L'indexation classique utilisée notamment en Recherche d'Information (RI), a pour but principal de représenter un texte (document, requête) par des mots-clés ou termes d'indexation (Salton G. & McGill M.J., 1983) (Baeza-yates R. et al., 1999) (Soulé-Dupuy C., 2001) (Calabretto S., 2003). L'indexation des textes comprend généralement deux étapes : la recherche des termes déterminant le contenu et l'évaluation du pouvoir de caractérisation de ces termes par des approches statistiques. Différents problèmes sont à résoudre :

- Définir l'élément qui sera choisi comme consentement d'indexation (radical, mot simple, groupe de mots).
- Distinguer les termes qui énoncent mieux un document et ceux qui ne le sont pas, en fonction du contenu du document (termes d'indexation).
- Juger du pouvoir de discrimination de ces termes : Certains termes sont plus importants que d'autres dans la détermination du contenu selon un test statistique fondé sur la fréquence d'apparition des termes.

Ainsi, plusieurs traitements morphosyntaxiques puis statistiques sont appliqués (Bruandet M-F et al, 1997) :

1. *Prétraitement et transformation* des textes : Une analyse morphologique est effectuée sur les textes ; elle consiste à les diviser en phrases selon des séparateurs de phrases fréquemment reconnus tels que les séparateurs de paragraphes (retour à la ligne) et les signes de ponctuations (le point, le point-virgule, etc.).

2. *Suppression des mots vides ou athématiques* en fonction de la langue du texte en utilisant un anti-dictionnaire.

3. *Radicalisation (stemming* en anglais*)* : Identifier à partir des différentes formes d'un terme (i.e. les différentes variantes morphologiques) l'unité d'indexation neutre simple.

4. *Pondération des termes* (mots simples ou mots-composés) pour différencier les termes qui n'ont pas le même pouvoir discriminant et ne représentent pas le contenu du document avec la même importance.

L'ensemble de ces traitements est décrit dans les sections suivantes.

1.4.1 Prétraitement et transformation des textes

Les données textuelles sont un aspect original de données complexes. Elles ne sont pas délimitées, structurées et étiquetées sémantiquement de façon explicite. En conséquence, ces

données demandent un traitement préalable permettant de minimiser l'espace de recherche. Cette phase est réalisée grâce à des techniques de prétraitement des textes comme le nettoyage des textes et l'élimination des caractères spéciaux peu informatifs. Nous nous intéressons aux approches basées sur un étiquetage morphosyntaxique.

- **Approche morphosyntaxique**

L'étiquetage morphosyntaxique convient à la préparation des textes pour la phase de modélisation du contenu. Il comprend une analyse morphologique et une analyse syntaxique (Chauché J., 1984), (Brill E., 1992). Notons que certains prétraitements (traitement des ponctuations, majuscules, codages et formats) précèdent ces deux analyses. En fonction des séparateurs de termes (espace, virgule, etc.) chaque phrase identifiée est décomposée en termes. Pour que ces termes deviennent des termes significatifs d'indexation (Fay-Varnier C., 1991), l'étiquetage morphosyntaxique associe à chacun sa catégorie morphologique (genre, nombre) et syntaxique (nom, adjectif, verbe, etc.).

Exemple :

Soit le document D1 = « Java est un langage de programmation objet. ». Une analyse morphosyntaxique de ce document produit : « Java – nom commun féminin singulier », « est – verbe indicatif présent 3ème personne du singulier », « un – article indéfini masculin singulier », « langage – nom commun masculin singulier », « de – préposition », « programmation – nom commun féminin singulier », « objet – nom commun masculin singulier », «. – ponctuation forte (fin de phrase) ».

Un traitement des caractères spéciaux détectés en cours de l'analyse morphologique est effectué après avoir affectée la catégorie à chaque mot du texte.

- **Traitement des caractères spéciaux**

Dans de nombreux systèmes d'indexation, un prétraitement du texte est réalisé afin de minimiser les variantes des termes et de simplifier la représentation des textes. Ces prétraitements consistent généralement à :

- Effacer les valeurs numériques : Dans les systèmes fondés sur des approches statistiques, les valeurs numériques sont difficiles à mettre à profit. Par exemple, pour un texte contenant « 35 millions d'euros », il faudrait pouvoir détecter qu'il s'agit d'une valeur correspondant à une somme d'argent exprimée en millions et offrir à l'utilisateur des opérateurs d'interrogation appropriés. Or, la plupart des systèmes de recherche actuels s'appuient sur les chaînes de caractères, mais pas sur le contenu sémantique.

- Supprimer les accents des caractères accentués : Cet suppression réduit les traitements de radicalisation et permet de borner les variantes des termes. En revanche, ce type de traitement mène à des ambiguïtés.
- Prétraiter les minuscules/majuscules : Souvent, une transformation en une forme unique est effectuée sur les lettres minuscules et majuscules (tout en minuscule). Cela évite des traitements du type « le terme apparaît après un point, donc il s'agit d'une majuscule de début de phrase ». En revanche, cela peut mener à des ambigüités (par exemple la non distinction des noms propres et des noms communs) qui ne pourraient être levées que par des traitements linguistiques (comme dans « M. Maison habite à Toulouse » où « Maison » sera considéré par un système d'indexation classique par le nom commun « maison » et non comme le nom d'une personne).

1.4.2 Utilisation d'un anti-dictionnaire (ou liste de mots vides)

Un anti-dictionnaire (*stop-list* en anglais) peut être utilisé pour éviter de garder des termes d'indexation qui ne conviennent pas à la sémantique du texte, c'est-à-dire qui ne correspondent pas aux thèmes traités dans le document.

Un anti-dictionnaire comporte généralement les mots vides (articles, pronoms, prépositions, locutions, adjectifs démonstratifs, relatifs et possessifs, verbes auxiliaires, mots outils, etc.) et les mots athématiques c'est-à-dire qui ne permettent pas de différencier les documents et se retrouvent dans n'importe quel texte indépendamment de son contenu.

Plusieurs anti-dictionnaires ont été produits et sont directement disponibles (en accès libre via le Web notamment), en particulier pour la langue anglaise. Les mots de ces anti-dictionnaires correspondent généralement à des termes qui ont une fréquence d'apparition élevée dans la collection de documents. La présence de ces termes dans la plupart des textes ne permet pas de discriminer, pour une requête, les textes pertinents de ceux qui ne le sont pas. En conséquence, ils ne sont pas intéressants pour l'indexation de textes.

1.4.3 Radicalisation

Un radical permet de représenter les différentes formes ou variantes morphologiques d'un mot par une même unité d'indexation. Il s'agit d'une forme neutre d'un mot : adjectif au masculin singulier, nom commun au singulier, verbe à l'infinitif, etc. Il existe plusieurs techniques de radicalisation qui se rassemblent en deux catégories : *techniques statistiques* et *techniques linguistiques* (Bouillon P. et al, 2000). Cette radication s'effectue :

- Soit par utilisation de glossaires qui proposent des radicaux (représentants de groupes) où chaque radical rassemble les différentes variantes morphologiques d'un mot. Par exemple, le nom masculin singulier pourra aussi être choisi comme représentant de

groupe, un verbe à l'infinitif pourra être le représentant de toutes les formes dérivées par conjugaison.

- Soit par élimination des suffixes par troncature droite des mots ; par exemple radicalisation à *x* caractères significatifs (Naffakhi Najeh M., 2013) selon une analyse statistique de la longueur des mots possédant des variantes morphologiques dans un même corpus de documents. Par exemple, la troncature droite des mots "francophone, francophonie, francophobe" est "francopho" (plus long radical commun) qui implique une troncature à 9 caractères pour ce groupe de termes.
- Soit par élimination des suffixes par utilisation d'un lexique qui comprend tous les suffixes possibles. Ce même principe peut être employé pour l'élimination des affixes. Des études statistiques sur les terminaisons des mots réalisent la création des lexiques de suffixes.
- Soit par radicalisation des mots en s'appuyant sur les règles de construction des mots (élimination des '*s*' terminaux, élimination des terminaisons comme '*tion*',...). On peut alors avoir des limitations sur les éliminations de suffixes comme, par exemple, ne pas éliminer le suffixe si le mot résultant a moins de quatre caractères ou éliminer le suffixe "-tion" sauf dans "nation", "lotion"... ou bien pas de restriction à l'élimination. Des règles de recodage ou d'association supplémentaires peuvent également apparaître (par exemple "chienne" qui devient "chienn" par élimination de la marque de féminin sera recodé en "chien").

Ces techniques sont appliquées pour toutes les langues. Il existe un algorithme de troncature et/ou de radicalisation pour chaque langue. A titre d'exemple, l'algorithme de *PORTER* (Porter, 1980), basé sur le principe de radicalisation par règles, est le plus utilisé pour la langue anglaise. Différentes versions de cet algorithme existent pour différentes langues, mais il ne peut être converti dans toutes les langues. Par exemple, en raison de variantes morphologiques et de règles de déclinaison plus complexes, il n'a pas d'équivalent en langue française. Dans ce cas, on a plutôt recours à des mécanismes de troncature droite (troncature variable).

1.4.4 Pondération des termes (mots simples ou mots-composés)

La pondération des termes est une mesure statistique qui peut être modifiée selon des critères différents (mots clés, domaines d'information, ...). En fait, le principe de pondération est basé sur l'hypothèse selon laquelle « *lorsqu'un auteur écrit un texte, il répète certains termes pour développer un aspect du sujet* » (Luhn, 1958).

La plupart des moteurs d'indexation utilisent des éléments statistiques pour pondérer l'importance d'un terme, notamment la *fréquence relative* et la *fréquence absolue* :

- La fréquence relative d'un terme dans un document correspond au nombre d'occurrences du terme dans le document.
- La fréquence absolue d'un terme correspond à la fréquence d'apparition de ce terme dans la collection de documents.

En conclusion, l'indexation classique, basée sur une étude morpho-lexicale et statistique de textes, ne traite pas réellement la sémantique proprement dite des documents. De ce fait, si on a besoin de prendre en compte le sens des termes extraits, il est nécessaire de s'orienter vers une indexation sémantique. Cette indexation sémantique possède de nombreuses perspectives : désambigüisation, reformulation automatique, meilleur ciblage thématique, etc.

1.5 Indexation sémantique

L'indexation sémantique représente les documents par des descripteurs qui reflètent plus précisément le contenu sémantique des documents qu'une simple liste de mots-clés pouvant souvent s'avérer ambigus (Aussenac G. N. et al., 2004). Cette indexation sémantique permet donc de décrire tout texte, d'une requête ou d'un document, par des descripteurs. La richesse sémantique des descripteurs utilisés influence la qualité de l'indexation, à savoir : les termes simples, les termes composés, les concepts et les relations inter-concepts. En effet, l'apport primordial de l'indexation sémantique est d'améliorer la représentation des documents en utilisant des termes d'enrichissement. Ces termes sont sémantiquement proches des termes d'indexation d'origine tels que les synonymes et les termes sémantiquement liés. Cependant, l'utilisation des représentations externes aux documents traités est obligatoire pour une indexation (Hernandez N., 2005) (Seydoux F., 2006). Ces descripteurs sont issus de ressources sémantiques externes telles que les taxonomies, réseaux sémantiques, thésaurus ou encore ontologies. L'identification des descripteurs associés à un document est basée sur l'utilisation d'une ressource sémantique. Traditionnellement, une ressource sémantique est constituée de termes, de concepts et de relations entre ces concepts.

Nous décrivons dans les paragraphes suivants les principaux types de ressources sémantiques que nous avons étudiés (les plus courants dans les travaux traitant d'indexation sémantique). Nous présentons ensuite les deux démarches principales sur lesquelles s'appuie l'indexation sémantique de documents, à savoir *l'identification des concepts*, et pon*dération des concepts*.

1.5.1 Différents types de ressources sémantiques

Dans l'indexation, il est souhaitable de prendre en compte le maximum d'informations concernant les descripteurs. Le jeu d'indexation utilisé est composé de descripteurs présents dans les documents (extraits de ces documents) et de descripteurs additionnels issus d'une

ressource sémantique. Nous présentons quatre types de ressources sémantiques, à savoir : les taxonomies, les thésaurus, les bases lexicales et les ontologies.

1.5.1.1 La taxonomie

La taxonomie[5], anglicisme de la notion de taxinomie (du grec *taxis* « placement », « classement », « mise en ordre » et *nomos* « loi », « règle ») fait référence à l'origine à la « Science des lois et des principes de la classification des organismes vivants, et par extension à la science de la classification »[6] (cf. Figure 4). Elle ne s'appliquait donc à l'origine qu'à la hiérarchie des organismes vivants en bactériologie, en botanique et en zoologie, mais son utilisation s'étend aujourd'hui à d'autres sciences, telles les sciences de l'information et les sciences humaines (Gilchrit A., 2010).

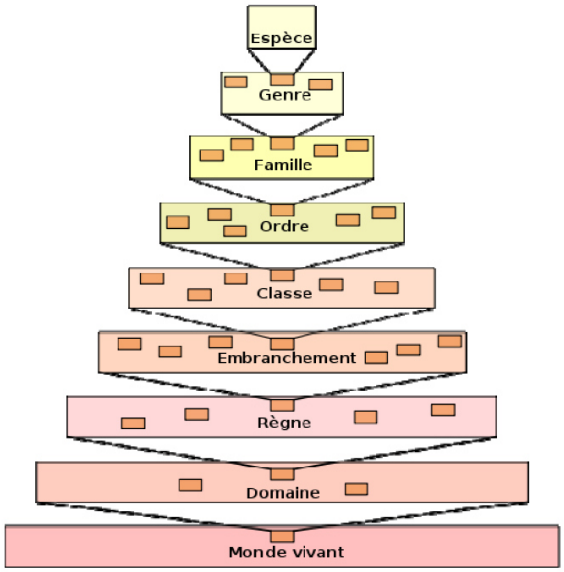


Figure 4. Hiérarchie taxonomique du vivant[7]

Une taxonomie traduit ainsi une « classification hiérarchisée d'objets observés et dénommés à l'intérieur d'une culture » représentée par des relations de subsomption, c'est-à-dire qu'un élément fait forcément partie d'un autre. Elle est schématisée par un arbre où chaque nœud est un taxon. Un taxon[8], synonyme de « groupe » en biologie, correspond à regroupement d'objets (ou éléments du vivant en biologie par exemple) ayant des caractères communs.

De nombreuses taxonomies sont disponibles à ce jour dans différents domaines. Il est alors important de pouvoir les cataloguer selon les domaines dont elles relèvent de sorte à pouvoir les utiliser dans des processus de classification, d'indexation et de caractérisation de ressources documentaires.

5 http://wiki.univ-paris5.fr/wiki/Taxinomie
6 http://www.cnrtl.fr/definition/taxonomie
7 Source : http://fr.wikipedia.org/wiki/Fichier:Taxonomic_hierarchy.svg
8 http://memsic.ccsd.cnrs.fr/mem_00575053/document

Cette organisation hiérarchique de concepts d'un domaine donné est partagée aujourd'hui par d'autres types de ressources terminologiques telles que les thésaurus et les ontologies. Ces ressources se différencient toutefois par leurs objectifs et leurs usages, selon que l'on souhaite décrire, indexer ou caractériser des contenus[9]. C'est ce que nous allons nous attacher à mettre en avant dans ce qui suit.

1.5.1.2 Le thésaurus

Un thésaurus est un glossaire contrôlé. Il rassemble un ensemble de termes structurés appelés descripteurs désignés pour leur tendance à décrire un domaine. Ces descripteurs sont exploités dans le but de décrire d'une manière précise le contenu des documents. Ils sont sélectionnés et normalisés pour l'indexation et le classement des documents. Dans un thésaurus, les termes représentent les concepts d'un domaine particulier. Des relations sémantiques relient ces concepts entre eux : liens hiérarchiques (généralisation et spécialisation), synonymie… (Gammoudi M., 1993). Afin de nommer facilement un terme descripteur, on lui associe un concept. Les termes d'un thésaurus peuvent aider à indexer des documents. En pratique, il existe des projets qui proposent des thésaurus pour l'indexation des documents, comme par exemple les projets MDweb[10] et NOESIS (Patriarche R. et al., 2005). Dans le domaine de l'environnement, MDweb est un outil de catégorisation d'informations géographiques qui utilise le thésaurus GEMET[11] pour indexer des documents. Dans le domaine des maladies cardiovasculaires, NOESIS est une plateforme d'aide au diagnostic médical dans laquelle les documents sont indexés en utilisant le thésaurus *UMLS*[12] "Unified Medical Language System" (Lindberg D. et al., 1983).

Dans le domaine médical, les thésaurus les plus utilisés sont les suivants :

- Le projet Unified Medical Language System (*UMLS*) lancé en 1986 par la NLM "National Library of Medicine". Il se compose de : (1) Un réseau sémantique qui se compose de 54 relations et 135 types sémantiques, et (2) Un méta-thésaurus défini dans le domaine biomédical, qui recueille des millions de termes appartenant à des terminologies et des nomenclatures ;
- Le thésaurus *MeSH* défini par National Library of Medicine (NLM, Bethesda, USA) englobe un vocabulaire contrôlé. Il est utilisé essentiellement dans le domaine du biomédical et de la santé. Sa principale fonction réside dans l'indexation et la recherche d'informations (et de documents). *MeSH* contient plus de 455 000 mots et plus de 25000 descripteurs.

[9] http://blog.sparna.fr/2013/12/07/ontologie-thesaurus-taxonomie-web-de-donnees/
[10] http://www.mdweb-project.org/
[11] http://www.eionet.europa.eu/gemet
[12] http://www.nlm.nih.gov/research/*UMLS*

1.5.1.3 La base lexicale

(Fellbaum C., 1998) définit une base lexicale ou un réseau sémantique comme étant « *un format de représentation permettant de mémoriser le sens des mots, pour rendre possible leur utilisation à la manière de l'être humain »*.

A ce jour, *WordNet*[13] est la base lexicale la plus connue. Cette base lexicale électronique est développée depuis 1985 par une équipe de psycholinguistes et de linguistes sous la direction de G. Miller à l'université de Princeton. *WordNet* qui est représenté par un réseau de nœuds et de liens, comprend la plupart des adjectifs, des noms, des verbes et des adverbes d'une langue donnée.

Chaque nœud, appelé *synset* (*syn*onym *set*), est constitué d'un ensemble de termes synonymes. Les termes synonymes sont rassemblés dans un nœud pour former un *synset*. Chaque *synset* représente un sens unique d'un mot particulier. Un terme peut être un mot simple ou une collocation (i.e. un terme complexe composé de deux mots ou plusieurs mots reliés).

Des liens ou des relations sémantiques relient les *synsets* de *WordNet*. La synonymie est la relation de base entre les termes d'un même *synset*. Les différents *synsets* sont quant à eux liés par diverses relations sémantiques telles que la relation de composition (méronymie-holonymie ou « partie-tout ») et la relation de subsomption (hyperonymie-hyponymie ou « est-un »).

WordNet a initialement été conçu autour de la langue anglaise, mais il existe aujourd'hui des initiatives dans différentes langues (notamment le français et autres langues européennes avec *EuroWordnet*).

Un des avantages, mais qui s'avère être aussi un des inconvénients majeurs des bases lexicales comme *WordNet*, c'est qu'elles sont très générales. En effet, n'ayant pas été formalisées pour modéliser un domaine donné, elles ne permettent pas de lever des ambiguïtés sémantiques dans le cas de termes polysémiques.

1.5.1.4 L'ontologie

La définition la plus citée présente une ontologie comme étant « *une spécification explicite et formelle d'une conceptualisation partagée »* (Gruber T.R.., 1993). Cette définition s'explique ainsi :

- *Explicite* signifie que « le type des concepts et les contraintes sur leurs utilisations sont explicitement définies » ;
- *Formelle* se réfère au fait que la spécification doit être lisible par une machine ;

[13] htt ://www.cogsci.princeton.edu/~

- *Conceptualisation* se réfère à « un modèle abstrait d'un certain phénomène du monde reposant sur l'identification des concepts pertinents de ce phénomène ».
- *Partagée* se rapporte à la notion selon laquelle une ontologie « capture la connaissance consensuelle, qui n'est pas propre à un individu mais validée par un groupe » ;

(Neches S. et al., 1991) définit aussi l'ontologie de la façon suivante : « *une ontologie définit les termes et relations de base constituant le vocabulaire d'un domaine, ainsi que les règles de combinaison des termes et des relations pour la définition d'extensions du vocabulaire »*.

On distingue deux types d'ontologies : les *ontologies légères* et les *ontologies lourdes*. Ces ontologies se différencient par la présence ou non d'axiomes (Aussenac G. N. et al., 2004). Les *axiomes* constituent des assertions, acceptées comme vraies, à propos des abstractions du domaine interprétées par l'ontologie.

Les ontologies légères sont composées seulement de concepts et de relations entre les concepts, elles sont dites « moins formelles ». Contrairement aux ontologies légères, les ontologies lourdes sont dites « formelles » (Ding Y. & Engels R., 2001), elles incorporent les règles d'inférence et les axiomes, en plus des concepts et des relations.

Les ontologies utilisées dans le domaine de la recherche d'information sont des ontologies légères. Elles se limitent à la définition des concepts et des relations entre les concepts. Les systèmes OntoQuery[14], Chemenet[15] et CIDOC/CRM[16] (Bruce Croft W. et al., 2008) constituent des exemples d'utilisation des ontologies en recherche d'information. Dans le domaine médical, les ontologies les plus utilisées sont : Gene Ontology[17] (GO), Galen[18] (système dédié à l'élaboration de l'ontologie dans tous les domaines médicaux).

1.5.1.5 Différences entre taxonomie, thésaurus, base lexicale et ontologie

Les principales différences entre les différentes ressources étudiées résident dans : 1) Les types de relations utilisées, 2) Le niveau de formalisme, 3) Le contenu, et 4) L'usage pour lequel elles ont été créées.

	Taxonomie	**Thésaurus**	**Base lexicale**	**Ontologie**
Types de relations utilisées	Hiérarchiques seulement	Hiérarchiques et associatives ; éventuellement relations d'alignement	Sémantiques : subsomption (hyperonymie-hyponymie : est-un), et la relation de	Inclusion (classe/sous-classe) : Opérations ensemblistes : union, intersection, exclusion.

[14] http://www.ontoquery.dk/index.php
[15] http://www.achemenet.com/
[16] http://cidoc.ics.forth.gr/
[17] http://www.geneontology.org/
[18] http://www.opengalen.org/themodel/ontology.html

			composition (méronymie-holonymie : partie-tout).	Caractéristiques des propriétés : transitivité, propriétés inverses, etc.
Niveau de formalisme logique	Moyennement formel	Peu formel	Peu formel	Très formel (formalisation mathématique)
Contenu	Des catégories organisés hiérarchiquement	Des concepts et des termes, organisés entre eux, avec leurs libellés, leurs traductions, leurs synonymes, et leurs descriptions/définitions.	Synset : (*synonym set*), un groupe de mots interchangeables, dénotant un sens ou un usage particulier	Des classes, des propriétés, et des règles logiques formelles. Eventuellement des instances de classe.
Utilisation	Sert à clasifier, à caractériser et à indexer des contenus ou des ressources	Sert à indexer des contenus ou des ressources avec des mots-clés et à les rechercher avec les mêmes mots-clés d'indexation	Sert à classifier, à caractériser et à indexer des contenus ou des ressources	Sert à instancier et à raisonner

Tableau 1. Comparaison entre taxonomie, thésaurus, base lexicale et ontologie[19].

Ces ressources sémantiques servent à analyser et indexer des textes plus ou moins longs. Le choix d'une ressource dépend essentiellement du type d'usage qui va en être fait et des apports attendus. Les bases lexicales, comme *Wordnet,* sont utilisées dans des contextes très généraux, comme la recherche d'informations sur le Web ou tout autre contexte multi-domaine. En revanche, lorsque l'on souhaite indexer des informations d'un domaine particulier, on préfère des ressources plus spécifiques (dites « de domaine ») qui vont permettre de désambigüiser les termes polysémiques, de généraliser ou de spécifier les propos, et ainsi de représenter de façon plus précise la sémantique du contenu des documents traités. Dans ce cas-là, des taxonomies ou des ontologies de domaine s'avèrent alors plus appropriées. Dans notre cas, pour la détermination de structure sémantique de documents XML orientés-textes, nous avons opté pour les taxonomies car nous nous intéressons particulièrement aux relations hiérarchiques entre les concepts.

Nous présentons dans ce qui suit les deux étapes nécessaires pour une indexation sémantique, à savoir : *identification* et *pondération* des concepts dans les documents, qui vont se baser sur l'usage d'une ressource sémantique de domaine, soit une taxonomie.

1.5.2 Identification des concepts dans les documents

Il existe deux approches d'identification de concepts dans des textes :

- Une première approche consiste à identifier et repérer, manuellement dans les contenus textuels des documents les concepts correspondant aux entrées d'une

[19] http://blog.sparna.fr/2013/12/07/ontologie-thesaurus-taxonomie-web-de-donnees/

ressource sémantique. Cette approche, suivie dans (Vallet D. et al., 2005), est généralement réalisée par un expert et a pour avantage d'être fiable car l'expert interprète la sémantique associée aux concepts dans la ressource et sélectionne le concept représentant au mieux la notion abordée dans le document. Cependant, même assisté par des traitements automatiques, cette manière reste coûteuse en temps, fastidieuse et sujette à des erreurs (Erdmann M.et al., 2000).

- D'autres approches visent à automatiser ce procédé. Cette automatisation dans le processus d'extraction des concepts est légitime dans la mesure où l'utilisation d'une ontologie permet d'accéder à la connaissance et à la rendre manipulable par les systèmes informatiques. Dans ce cas, les labels ou termes désignant les concepts sont recherchés dans les documents. Un concept est en effet défini à partir d'un ou plusieurs labels représentant les variantes lexicales que peuvent prendre les termes définissant les concepts (Vallet D. et al., 2005) (Kiryakov A. et al., 2004).

A partir d'une ressource terminologique, cette identification de concepts et d'instances dans les documents se base sur trois étapes :

- *Extraction des termes du document* : Elle consiste à extraire l'ensemble des termes apparaissant dans les documents.
- *Recherche des labels correspondant à des concepts* : Les labels sont recherchés dans l'ensemble des termes extraits en favorisant la prise en compte des labels les plus longs et donc des concepts les plus spécifiques (Baziz M., 2005), (Vallet D. et al., 2005).
- *Désambiguïsation des labels* : Les labels peuvent cependant se rapporter à plusieurs concepts. Dans ce cas, et afin d'identifier quel est le concept abordé dans le document, un mécanisme de désambiguïsation du terme est mis en place la plupart du temps identification du domaine thématique du texte en question (Baziz M., 2005).

1.5.3 Pondération des concepts extraits d'un document

Le calcul du poids d'un concept dans la représentation d'un document peut être effectué selon deux catégories de méthodes : statistiques ou sémantiques.

- *Pondération statistique* : La pondération statistique se base sur une adaptation de la mesure *TF-IDF* (Baeza-Yates R. et al.,1994) (Soulé-Dupuy C., 2001) où TF (pour "Term Frequency") est la *fréquence relative* du terme dans le document, et IDF (pour "Inverse Document Frequency") est une *fonction inverse* de la *fréquence absolue* du terme dans la collection de documents. Cette mesure est utilisée en *Recherche d'Information* (RI) pour calculer le pouvoir discriminant d'un terme. Appliquée aux concepts, cette mesure vise à défendre les concepts apparaissant fréquemment dans un

document mais faiblement dans le reste de la collection de documents (Hermandez N., 2006). Cette pondération est similaire à celle présentée dans (Baziz M., 2005) et (Vallet et al., 2005). Cependant, une partie de la sémantique contenue dans les relations entre concepts est ignorée, seule la notion statistique de co-occurrence est prise en compte.

– *Pondération sémantique :* La pondération sémantique prend en compte le lien dans la ressource sémantique entre le concept considéré et les autres concepts du document et illustre le contexte sémantique du concept. Elle repose sur le principe suivant : *Plus un concept est proche sémantiquement des autres concepts retrouvés, plus il est représentatif de l'ensemble des thématiques du document* (Hermandez N., 2006). La représentativité sémantique d'un concept est calculée à partir de la proximité du concept considéré avec les autres concepts retrouvés dans le document.

Il existe des outils pour l'extraction des concepts à partir des documents ; dans ce qui suit nous nous intéressons aux principaux outils du domaine.

1.5.4 Principaux outils d'extraction de concepts

Nous présentons dans cette section les outils d'extraction de concepts accessibles en ligne ou téléchargeables, à savoir : *PubMed ATM*[20], *MetaMap* (Aronson A. R., 2001), *MTI (Aronson A. R. et al., 2004)* et *MaxMatcher* (Zhou X. et al., 2006).

1.5.4.1 PubMed ATM

PubMed ATM est un service implanté dans le portail de *PubMed* visant à associer un morceau de texte (par exemple, la requête de l'utilisateur) à des termes ou concepts dans les différentes tables et les index associés dans l'ordre suivant[21] : 1) Table des termes désignant les concepts *MeSH* ainsi que les informations supplémentaires comme qualificatifs, types de publication, substances ... ; 2) Table des Journaux ; et 3) Table des Auteurs. Étant donnée une requête, *PubMed* essaie de localiser les groupes de mots les plus longs qui sont sauvegardés dans les tables de concepts. Lorsqu'un terme désignant un concept est trouvé, le processus de recherche du terme candidat est terminé.

Ensuite, les termes retrouvés sont groupés par des expressions booléennes pour reformuler une requête booléenne. Si aucun terme n'est trouvé dans les tables, les mots sont combinés par l'opérateur "AND" pour rechercher des documents contenant tous les mots.

La stratégie d'extraction des concepts du service *ATM* de *PubMed* repose sur une recherche exacte des termes dans la base de données. Il est capable de retrouver facilement les termes synonymes ainsi que les variantes d'un terme donné. Par contre, le problème suivant peut être observé lors de l'extraction des concepts : *PubMed ATM* essaie plusieurs

[20]http://www.ncbi.nlm.nih.gov/pubmed
[21]http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.How_PubMed_works_aut.

combinaisons possibles des mots pour formuler une nouvelle requête booléenne qui la rend finalement plus compliquée et plus difficile à interpréter par l'utilisateur.

1.5.4.2 MetaMap

MetaMap est un programme permettant d'extraire les concepts du thésaurus *UMLS* dans des textes biomédicaux (Aronson A. R., 2001). Il s'agit du composant principal de l'outil *MTI* qui est utilisé pour indexer les documents dans la base *MEDLINE*[22] (*Medical Literature Analysis and Retrieval System Online*), base de données bibliographiques[23] gérée par la NLM. Le processus d'extraction consiste à :

1. *Identifier des groupes nominaux* (lallich-Boidin et al., 2005) dans le texte à l'aide de l'utilisation de l'analyseur grammatical de Xerox (Cutting D. et al., 1992).
2. *Générer des variantes* (synonymes, acronymes, ...) pour chaque groupe nominal en utilisant la ressource *SPECIALIST Lexicon* d'*UMLS*.
3. *Sélectionner des concepts candidats* : Tous les concepts ayant au moins un mot qui se trouve dans une des variantes sont récupérés.
4. *Évaluer les concepts* : Les concepts candidats sont comparés avec le texte original à l'aide des quatre mesures suivantes : centralité, variation, couverture et cohérence (Aronson A. R., 2001). Les deux dernières dépendent de la fréquence d'apparition des concepts dans le texte. Les concepts candidats sont finalement ordonnés en fonction du score final.
5. *Construire les mappings* : pour chaque groupe nominal, les concepts pertinents sont sélectionnés en fonction d'un score de similarité.

MetaMap présente des avantages et des inconvénients rapportés par plusieurs chercheurs spécialistes du domaine. Trois inconvénients majeurs, qui influencent les performances de *MetaMap*, ont été identifiés :

- Le premier est lié au fait que la sélection des concepts candidats est basée sur les mots simples, ce qui pose un problème de sur-génération ("*over-generation*") des variantes liées aux concepts non-pertinents retournés. Par exemple, étant donné le groupe nominal "*ocular complications*", *MetaMap* l'associe à trois concepts "*Ocular*", "*Complications*" et "*Complications Specific to Antepartum or Postpartum*" car ils partagent au moins un mot en commun.
- Le deuxième inconvénient concerne la comparaison stricte entre chaque groupe nominal et les entrées dans le méta thésaurus. Cela pose un problème de sous-

[22]http://www.ncbi.nlm.nih.gov/pubmed?term=medline[sb]
[23]MEDLINE est une base de données bibliographiques regroupant la littérature relative aux sciences biologiques et biomédicales. La base est gérée et mise à jour par la bibliothèque américaine de médecine (NLM).

génération de variantes pertinentes. Par exemple, pour l'expression "*gyrb and p53 protein*", *MetaMap* n'arrive pas à identifier "*gyrb*" comme une protéine car celle-ci est enregistrée comme "*gyrb protein*" dans le thésaurus *MeSH* ou dans l'*UMLS*.

- Un autre inconvénient de *MetaMap* concerne le coût important en terme de temps de traitement parce que cet outil comporte un ensemble de méthodes linguistiques sophistiquées comme l'analyse grammaticale, la génération des variantes, la recherche dans l'ensemble du méta thésaurus, ainsi que le calcul de plusieurs mesures statistiques.

1.5.4.3 MTI "Medical Text Indexer"

MTI a été présenté dans (Aronson A. R. et al., 2004) comme outil d'assistance à l'indexation des citations d'articles de journaux dans *MEDLINE*. Plus précisément, *MTI* suggère les concepts les plus similaires pour chaque document aux indexeurs humains à la NLM. *MTI* est essentiellement basé sur les trois algorithmes suivants d'extraction de concepts : *MetaMap* (Aronson A. R., 2001), "*Related-PubMed citations*" (*PRC*) (Wilbur J., 2003) et "Restrict to *MeSH*" (Bodenreider O. et al., 1998). Le premier sert à localiser les concepts candidats potentiels tandis que le deuxième permet de pondérer les mots selon un schéma *TF-IDF (*cf. paragraphe 1.5.3), appelé selon les auteurs « poids local » et « poids global » des mots, en favorisant ceux qui apparaissent dans les documents voisins ou similaires à un document particulier. Le dernier algorithme a pour objectif de transformer les concepts issus du thésaurus en concepts *MeSH* pour mettre en évidence les sujets sémantiques du document.

La liste des concepts candidats peut contenir des concepts non-pertinents par rapport à la sémantique exacte du texte (Aronson A.R. et al., 2004). Pour pallier ce problème, *MTI* applique trois différents niveaux de filtrage pour éliminer les concepts non-pertinents :

- *Filtrage strict* : Consiste à supprimer tous les concepts candidats non-pertinents, c'est-à-dire ceux qui ne sont identifiés ni par *MetaMap*, ni par "*PubMed Related Citation*". Toutefois, il existe le risque que certains concepts pertinents soient éliminés.
- *Filtrage moyen* : Ce filtrage a pour objectif d'augmenter la mesure *rappel*[24] en assouplissant les conditions de sélection des concepts candidats.
- *Filtrage de base* : Par défaut, *MTI* utilise le filtrage basique qui est composé des trois fonctions suivantes : 1) Addition et suppressio*n des termes* désignant les conc*epts MeSH* ou les *qualificatifs*[25] en se basant sur les résultats de chacune des deux

[24]Le rappel est défini par le nombre de documents pertinents retrouvés au regard du nombre de documents pertinents que possède la base de documents. Cela signifie que lorsque l'utilisateur interroge la base il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre la requête de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé.

[25]Les qualificatifs représentent des concepts généraux employés pour préciser le sens des mots clés. L' affiliation d'un qualificatif à un mot clé est destinée à mettre en évidence un aspect particulier du concept représenté par ce mot clé et en accroît ainsi la spécificité.

méthodes utilisées dans les deux outils : *MetaMap* et *MTI* ; 2) Désignation des concepts identifiés par les deux méthodes ; et 3) Substitution des *qualificatifs* pour certains *concepts* identifiés comme candidats. En général, le filtrage de base renvoie plus de concepts candidats et par conséquent plus de concepts dans la liste finale.

1.5.4.4 MaxMatcher (Extractor)

MaxMatcher (Zhou X. et al., 2006) est un outil d'extraction de concepts basé sur la recherche de chaînes de caractères ou de termes composés de mots stockés dans un dictionnaire des concepts. La recherche d'une chaîne de caractères dans un dictionnaire de concepts peut se faire de manière exacte et/ou approximative. *MaxMatcher* utilise *MeSH* et *UMLS* comme ressources terminologiques pour identifier les concepts. Du fait que le thésaurus MeSH est maintenu par une organisation (la NLM), ceci ne contient pas de termes ambigus désignant les concepts. Par conséquent, MaxMatcher adopte la recherche exacte pour localiser les concepts si MeSH est utilisé. Par contre, pour l'UMLS, un terme peut désigner plusieurs concepts, ce qui représente l'ambiguïté dans l'UMLS (Dinh D., 2012).

Étant donné un texte, *MaxMatcher* la découpe en phrases puis localise la chaîne la plus longue qui correspond à une entrée dans le thésaurus *MeSH*. Pour la recherche approximative, la chaîne peut être plus courte que l'entrée du concept.

Par exemple, le mot ″gyrb″ (nom d'une protéine) est enregistré en tant que concept ″gyrb protein″ dans *UMLS*. Si un concept est constitué de deux sous-concepts, *MaxMatcher* retourne deux sous-concepts candidats plutôt que de retourner le concept le plus spécifique dans la hiérarchie.

Par exemple, l'expression ″**Ablation of liver tumor** by **injection** of **hypertonic saline**″ contient trois concepts mis en gras dont les identifiants dans *MeSH* sont respectivement : "**Ablation of liver tumor***″(C2004650,E04.014)*, "**injection***″(C0021485,* E02.319.267.530*)* et "**hypertonic saline**" *(C0036085,* D26.776.314.890*)*. Cependant, *MaxMatcher* retourne quatre concepts : ″Ablation″ (C0547070, T169) of ″liver tumor″ (C0023903, T191) by ″injection″ (C0021485, T061) of ″hypertonicsaline″ (C0036085, T121, T197).

1.5.4.5 Comparaison des outils d'extraction de concepts

Le tableau 2 compare les différents outils d'extraction de concepts présentés précédemment. Les lignes représentent les travaux et les colonnes correspondent aux critères de comparaison notés :

C1 : Type de recherche : Exacte ou approximative.

C2 : Extraction des concepts basée sur des termes simples.

C3 : Suppression des concepts non pertinents.

C4 : Extraction des concepts sémantiquement liés à un terme donné.

C5 : Ressources sémantiques par outil d'extraction.

Critères / Outils	**C1 :** Type de recherche	**C2 :** Extraction des concepts basée sur des termes simples	**C3** : Suppression des concepts non pertinents	**C4** : Extraction des concepts sémantiquement liés à un terme donné	**C5** : Ressources sémantiques par outil d'extraction
PubMed ATM	Exacte	Non	Non	Non	Une seule (*MeSH*)
MetaMap	Exacte	Non	Non	Non	Une seule (*UMLS*)
MTI	Exacte	Non	Oui	Non	Une seule (*MeSH*)
MaxMatcher (Extractor)	Exacte ou approximative	Oui	Oui	Oui	Deux (*MeSH* + *UMLS*)

Tableau 2. Comparaison des outils d'extraction de concepts.

Nous remarquons que, parmi tous ces outils, l'outil que nous jugeons le plus complet est *Extractor* car il est capable d'effectuer des recherches exactes ou approximatives selon le type de ressource utilisée, *UMLS* ou *MeSH* ; c'est à dire que *MaxMatcher* (*Extractor*) est l'outil le plus flexible en termes de type de recherche utilisée. De plus, son temps de traitement est moins important que les autres outils d'extraction des concepts à partir de textes (Dinh D., 2012).

Jusqu'à maintenant, nous avons présenté les concepts de base d'une indexation de façon générale à savoir : classique ou sémantique. Cette présentation sera complétée par l'état de l'art des travaux d'indexation et structuration sémantique de document (cf. du chapitre 2).

1.6 Conclusion

Ce premier chapitre nous a permis d'introduire et de définir l'ensemble des notions sur lesquelles reposent nos travaux de thèse. Dans une première partie, nous avons présenté les concepts de base concernant les documents *XML*, leurs différentes structures et catégories. La deuxième partie a été consacrée au concept d'entrepôt de documents (définitions et objectifs). Comme la problématique abordée dans cette thèse concerne principalement la structuration sémantique des documents, dans un contexte d'entreposage de documents, nous avons alors présenté par la suite les concepts de base d'un processus d'indexation : *Indexations classique* et *sémantique*.

La qualité de l'indexation sémantique dépend de la richesse de la ressource sémantique utilisée. Cette indexation sémantique repose sur deux étapes : 1) L'identification des concepts et instances dans les documents, et leur 2) Pondération.

Concernant l'étape d'identification des concepts, nous avons étudié les principaux outils d'extraction des concepts à partir de documents textuels et nous avons retenu l'outil *MaxMatcher* (*Extractor*) puisqu'il recherche de manière plus flexible, dans une ressource sémantique, les concepts qui correspondent aux termes extraits d'un document donné.

Pour ce qui est de la pondération, une méthode de pondération des concepts des taxonomies sera proposée et utilisée dans le processus de détermination de la structure sémantique d'un document (cf. chapitre 3) (Ben Meftah S. et al., 2012).

Dans le chapitre suivant, nous présentons les travaux abordant de façon plus approfondie l'indexation sémantique. Nous classifions ces travaux selon le type de ressource sémantique utilisée : *Spécifique* ou *Générale*.

Chapitre 2

Etat de l'art : Indexation et structuration sémantique des documents XML

Sommaire

2.1 Introduction

L'indexation sémantique vise à s'appuyer sur des ressources terminologiques sémantiques pour représenter la sémantique des documents. Elle repose sur l'intuition selon laquelle le sens des informations textuelles dépend des relations conceptuelles entre les objets du monde auxquels elles font référence plutôt que des relations contextuelles trouvées dans leur contenu (Haav H.M. & Lubi T.L., 2001). L'indexation sémantique n'est rendu possible que grâce à l'existence et l'utilisation de ressources sémantiques décrivant explicitement l'information correspondant aux objets (du monde qu'elles décrivent). Comme introduit dans le chapitre précédent, cette indexation repose sur deux étapes : 1) L'identification des concepts contenus dans les documents indexés, et 2) La pondération de ces concepts. Les ressources sémantiques utilisées sont classées en deux types : *Générales* (e.g., *WordNet*) et *Spécialisées* (*MeSH*, *UMLS*, *SOMED-CT*, *Menelas*...). Dans ce chapitre, nous présentons différents travaux de la littérature relatifs à l'indexation sémantique selon ces deux types de ressource sémantique.

2.2 Travaux utilisant une ressource sémantique générale

Dans ce qui suit, nous étudions différents travaux de l'état de l'art traitant d'indexation sémantique en utilisant une ressource sémantique générale.

2.2.1 Travaux de (Zargayouna H. et al., 2004)

Les auteurs proposent une méthode d'évaluation de similarité entre les concepts d'une ressource sémantique, mise en œuvre dans un système d'indexation sémantique de documents *XML*. Ainsi, le contenu textuel (des nœuds feuilles de l'arbre *XML*) est indexé en utilisant une ressource sémantique générale, *WordNet* en l'occurrence.

Les documents sont représentés sous forme de vecteurs de termes selon le modèle vectoriel de Salton (Salton G.et McGill M.J., 1983), tout en reliant ces termes aux concepts de la ressource sémantique *WordNet* utilisée lors de l'indexation sémantique. Les termes du document utilisateur seront par la suite remplacés par les concepts associés afin que les réponses soient plus appropriées aux utilisateurs. Le système d'indexation de (Zargayouna H. et al., 2004) permet l'exploitation à la fois de la structure et du contenu textuel des documents. La Figure 5 schématise ce processus d'indexation.

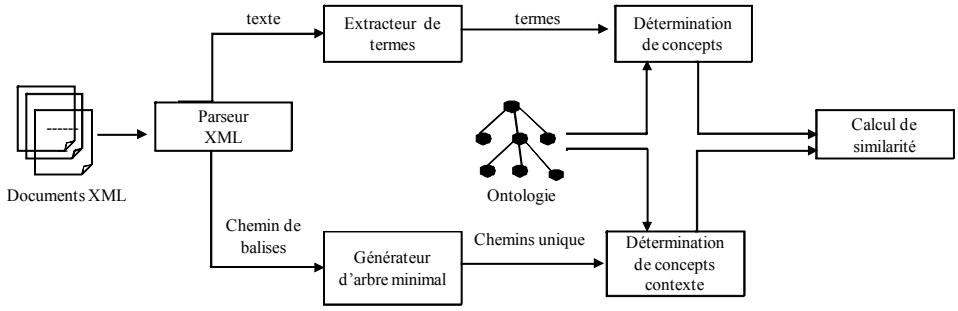


Figure 5. Processus d'indexation de (Zargayouna H. et al., 2004)

Le système proposé par Zargayouna possède les limites suivantes :

- L'ontologie est créée à partir du corpus, manuellement ou par des méthodes semi-automatiques. Le lien entre les termes et le concept est alors évident. Il faut noter que le problème d'appariement entre termes et concepts est plus complexe à établir lorsqu'on dispose d'un corpus de spécialité et d'une ontologie de domaine pré-définie (i.e. non spécifique au corpus).
- Une autre limite du système d'indexation proposé est la restriction au lien de spécification/généralisation (is-a) dans l'ontologie.

2.2.2 Travaux de (Kang B. Y. et Lee S., 2005)

Dans l'approche décrite dans (Kang B. Y. et Lee S., 2005), les auteurs ont proposé de considérer non seulement les termes mais aussi les concepts d'un document. Dans cette approche, les concepts d'un document sont extraits puis, à partir de ces concepts, sera dérivé un vecteur sémantique de concepts pondérés. Cette approche est illustrée dans la Figure 6.

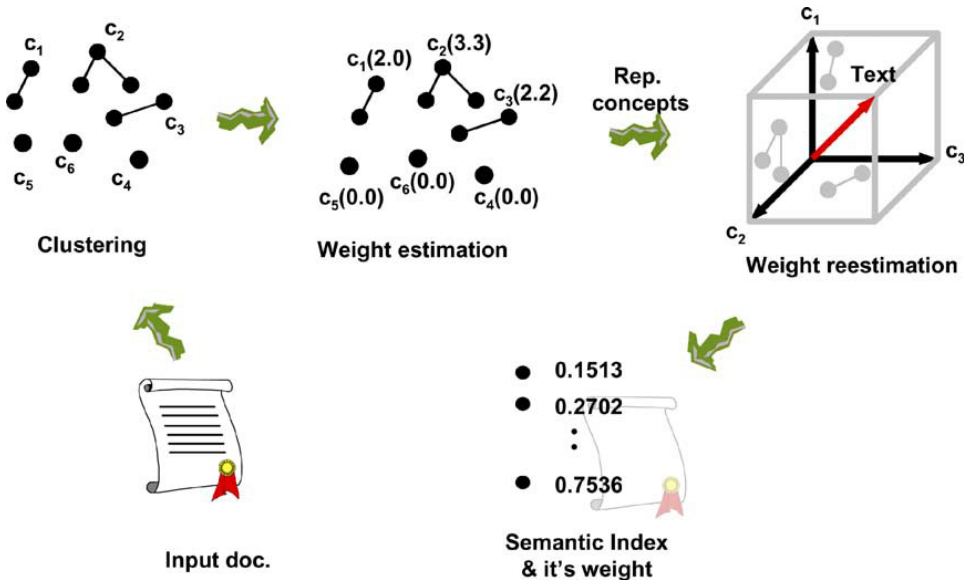


Figure 6. Flux du traitement sémantique (Kang B. Y. et Lee S., 2005).

L'approche proposée comprend trois étapes principales :

- Étape 1 : Extraction des chaînes lexicales (ou *clusters* dans la Figure 6) à partir du contenu des documents. Le « *clustering* » des chaînes lexicales permet d'identifier les concepts correspondants.
- Étape 2 : Estimation des poids des termes et des concepts (correspondants aux clusters de chaînes lexicales) sur la base des relations entre les termes composant chaque chaîne lexicale.
- Étape 3 : Re-calcul des poids des chaînes lexicales en se basant sur le poids des concepts.

La première étape s'appuie sur les quatre types de relations extraites de *WordNet* entre les concepts : identité, synonymie, hyperonymie et méronymie (partie de de). Cette étape est composée des trois sous étapes initialement proposées par (Kang B. Y. et Lee S., 2005) :(i) tout d'abord, sélection des différentes chaînes lexicales existantes dans le document ; ensuite, (ii) détection de la relation d'identité entre ces chaînes lexicales ; et enfin, (iii) classement des chaînes en fonction de leurs relations avec d'autres chaînes lexicales.

La deuxième étape permet le calcul des poids des termes et des concepts en se basant sur les relations entre les termes. Les termes ayant plus de relations avec d'autres termes sont considérés sémantiquement plus importants. Le même principe s'applique pour le calcul des poids des concepts extraits.

Après l'extraction des concepts représentatifs du document (ceux qui ont un score maximum), la troisième étape consiste à recalculer l'importance sémantique des termes car (Kang B. Y. et Lee S., 2005) ont constaté que le degré d'importance sémantique d'un terme est affecté par le poids du concept dans lequel il est inclus. Pour ce re-calcul, ils ont proposé un nouveau modèle vectoriel pour les chaînes lexicales extraites du document en exploitant le modèle vectoriel des concepts candidats construit à l'étape 2. Dans ce modèle, le document est représenté par un ensemble de concepts et le degré d'importance sémantique des chaînes lexicales est recalculé en tenant compte du poids du concept d'appartenance.

Quand deux chaînes lexicales ont le même score, la chaîne lexicale dont le groupe de concepts possède le score le plus élevé est considérée sémantiquement comme la plus importante des deux chaînes. Par exemple, dans la Figure 7, les deux termes t_1= "*practice*" du concept C4 et t_2="*director*" du concept C3 ont le même score 0.3, le concept *C4* contenant t_1 est plus important que le concept *C3* contenant t_2. Alors le terme t_1 est sémantiquement plus important que le terme t_2.

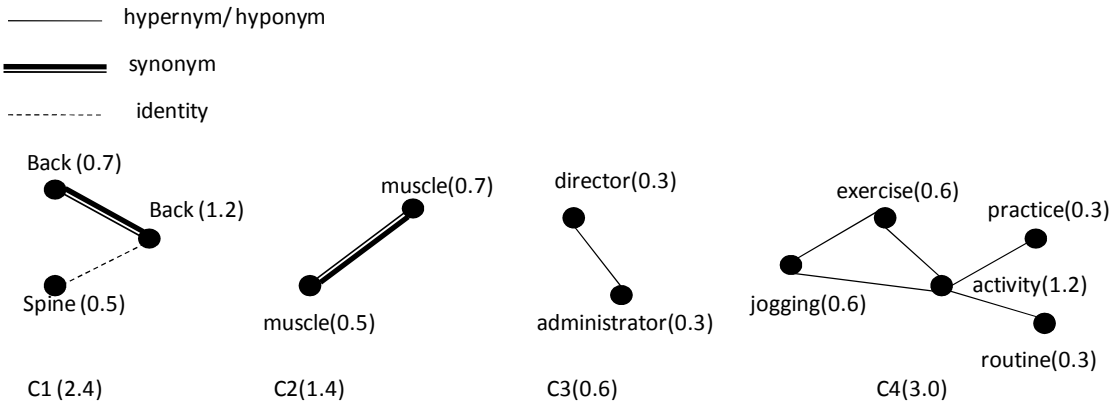


Figure 7. Importance sémantique des chaînes lexicales (Kang B. Y. et Lee S., 2005).

Ce travail présente deux avantages :

- Premièrement, l'utilisation de *WordNet* dans l'identification des concepts permet de simplifier cette étape. Cette simplification est due à ce que le nombre de catégories syntaxiques utilisées dans *WordNet* pour l'extraction des chaînes lexicales est limité par rapport à d'autres thésaurus.
- Deuxièmement, l'utilisation des relations sémantiques entre les noms des termes afin d'identifier les concepts permet d'améliorer les performances en RI.

Cependant, l'utilisation de la base lexicale générale *WordNet* pose un problème de couverture terminologique puisque seuls les termes des documents les plus communs sont référencés par des concepts. Ce qui implique que les termes spécifiques (de domaine) ne sont pas référencés.

2.2.3 Travaux de (Baziz M. et al., 2007)

Dans (Baziz M. et al., 2007), les auteurs proposent un modèle de représentation sémantique des documents par un réseau sémantique (ensemble de concepts reliés par des liens). Cette approche est composée de trois étapes (cf. Figure 8) :

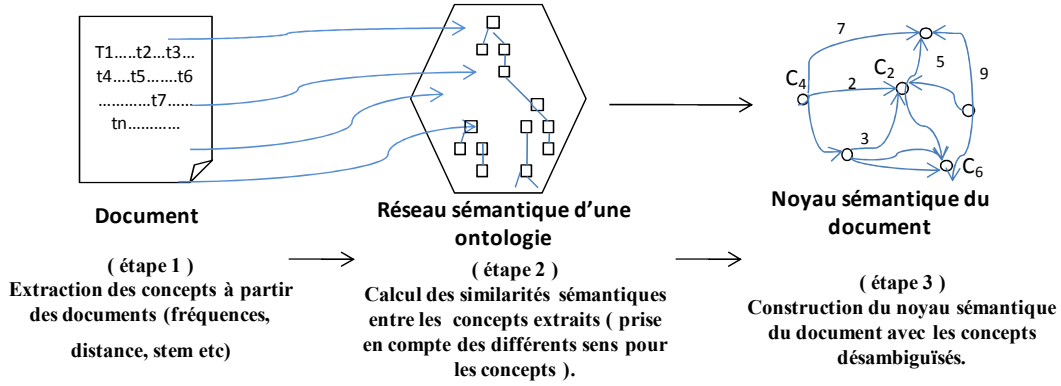


Figure 8. Schéma général de l'approche (Baziz M. et al., 2007).

L'étape 1 extrait les concepts les plus fréquents qui correspondent à des entrées dans l'ontologie à partir des documents.

Dans l'étape 2, les auteurs se servent de l'ontologie pour récupérer les différents sens possibles des concepts retenus. Ils calculent les similarités entre les différents sens de ces concepts en se basant sur les relations sémantiques disponibles telles que la généralisation/spécialisation, les liens de composition, et les liens de domaine.

L'étape 3 concerne la construction du réseau sémantique en choisissant le concept ayant le score le plus élevé qui permet de désambiguïser un concept donné retenu dans l'étape 2. Le concept ayant le score le plus élevé représentera un nœud dans le réseau sémantique.

Pour l'expérimentation du modèle, trois cas de figures ont été évalués :

- Cas 1. Indexation basée sur les mots clés : dans cette méthode classique, les poids des mots clés extraits des documents sont calculés suivant la formule *TF-IDF* classique.
- Cas 2. Indexation basée sur les concepts : seuls les nœuds (concepts-sens) des réseaux sémantiques sont utilisés et les poids des mots clés sont calculés suivant la formule *TF-IDF* pour les documents et les requêtes.
- Cas 3. Indexation basée sur les concepts et les mots clés : Les deux précédentes méthodes sont combinées afin de réaliser ce type d'indexation.

Cependant, leurs expérimentations ont montré que l'indexation sémantique (i.e., l'affectation d'un ensemble de concepts à un document) n'améliore les résultats que lorsqu'elle est combinée avec une indexation classique basée sur les mots-clés.

2.2.4 Travaux de (Tagarelli A. & Grec S., 2010)

(Tagarelli A. & Grec S., 2010) proposent une méthode d'enrichissement sémantique des noms des balises d'un document *XML*. Pour cela, chaque chemin[26] du document *XML* représente un réseau et chaque balise de ce chemin constitue une couche, contenant l'ensemble des sens de la balise en question extraits de *WordNet*. L'étape suivante consiste à calculer la mesure de similarité entre les différentes couches afin de trouver le meilleur chemin dans le réseau.

Cet enrichissement est analysé selon la structure ainsi que les noms des balises :

- Analyse de la structure s'applique aux chemins des balises *XML* afin d'éviter l'ambiguïté de la signification des noms des balises, exprimée en synonymie et polysémie.
- Analyse des balises textuelles en calculant la pertinence des termes des balises par la combinaison de la pertinence syntaxique et sémantique d'un terme. La

[26] L'ensemble des balises reliant la racine à un élément feuille du document.

pertinence syntaxique tient compte du contexte structurel du terme, tandis que la pertinence sémantique dépend du degré de polysémie d'un terme.

Dans ces travaux, la méthode d'enrichissement sémantique se compose de 6 étapes (cf. Figure 9) :

Étape 1 : Initialement, chaque document *XML* dans la collection est décomposé en un ensemble d'arbres.

Étape 2 : Les arbres de tous les documents sont rassemblés, et les caractéristiques de structure et de contenu sont générées à l'aide du thésaurus général *Wordnet*.

Étape 3 : Les arbres sont modélisés de façons que les éléments sont conçus pour intégrer la structure sémantique enrichie à partir des documents originaux.

Étape 4 : Les documents *XML* modélisés comme des transactions deviennent l'entrée d'un algorithme de classification.

Étape 5 : La tâche de classification donne une série de groupes de transactions *XML*. Cette classification peut être utilisée pour préparer une classification des documents *XML* originaux, afin de fournir à l'utilisateur une organisation des textes d'entrée qui est probablement plus utile.

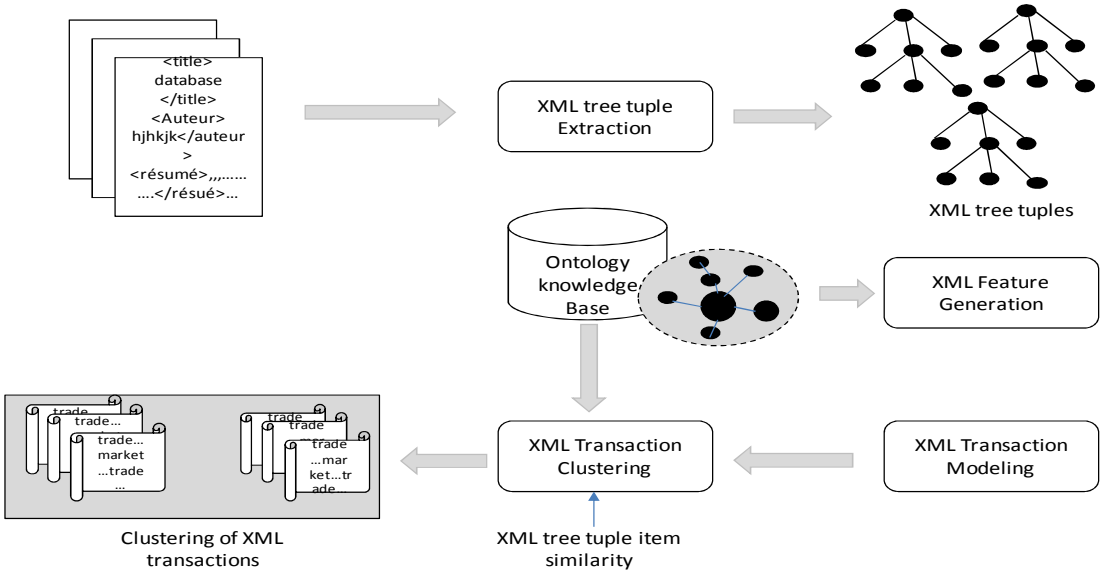


Figure 9. Approche d'enrichissement sémantique des documents XML (Tagarelli A. & Grec S., 2010).

Nous pouvons constater que l'utilisation de *WordNet* peut causer des difficultés pour le choix du sens le plus approprié pour une balise donnée surtout pour les termes polysémiques[27]. Ce qui représente une difficulté pour l'utilisateur lors de la recherche du document pertinent pour son besoin à cause des différentes significations des noms des balises.

[27]Un terme polysémique est un terme qui peut avoir plusieurs sens différents.

2.2.5 Travaux de (Egozi O. et al., 2011)

Dans (Egozi O. et al., 2011) est proposée une approche de RI à base de concepts en utilisant la ressource sémantique *WordNet* et la méthode d'Analyse Sémantique Explicite (*ESA*)[28]. Cette méthode d'analyse de texte est automatique, elle extrait les concepts représentant un document ou une requête.

- **Indexation de documents et de requêtes par concepts selon la méthode *ESA***

Les auteurs utilisent la méthode *ESA* (*Explicit Semantic Analysis*) pour transformer chaque document d'un corpus en un vecteur pondéré de concepts. Le poids des concepts représente le degré d'association entre le terme et le concept. Ce poids est calculé de la même façon que le poids d'un terme dans une indexation classique (pondération de type *TF-IDF*, cf. 1.5.3).

La méthode *ESA* n'associe qu'un nombre limité de concepts aux textes indexés. De fait, un problème de représentativité émerge lorsque les textes sont trop longs pour ce type de méthode. Afin de résoudre ce problème, les auteurs ont choisi de décomposer le document en segments de textes et de calculer un score pour chacun. Ensuite, ils classifient les segments selon leurs pertinences, du plus pertinent vers le moins pertinent.

Ce processus d'indexation est illustré dans la Figure 10. Par exemple, les concepts associés au terme *market* sont *Bazaar*, *NASDAQ, Economy* possédant comme degré d'association (poids) respectivement 0.72, 0.65 et 0.53. Ces concepts sont classés selon leurs pertinences du plus important vers le moins important.

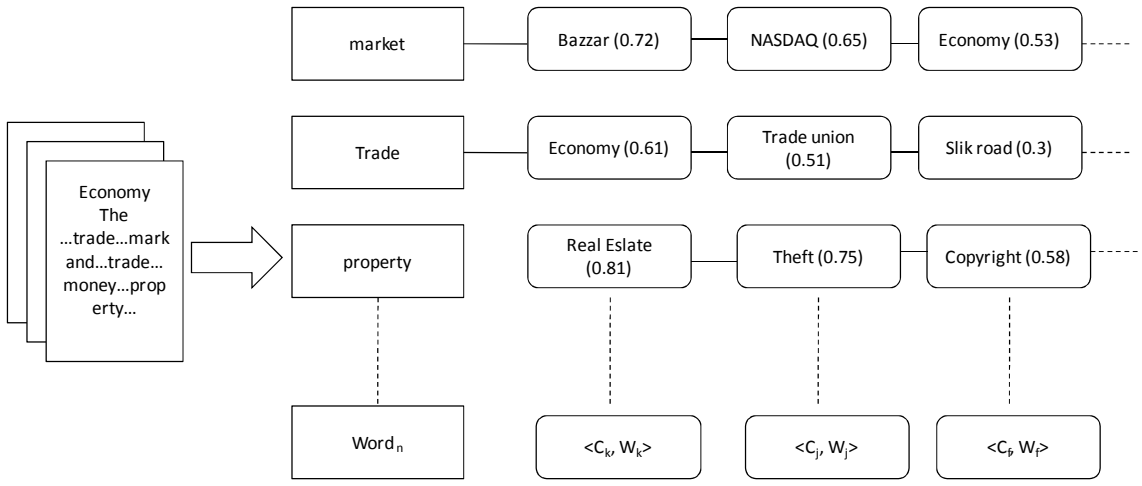


Figure 10. Génération d'un ESA d'un article Wikipédia (Egozi O. et al., 2011).

La méthode d'indexation proposée dans ces travaux s'applique aussi bien aux documents, qu'aux requêtes utilisateurs. Les requêtes sont donc indexées selon la même méthode que les documents, en utilisant *ESA* pour l'identification des concepts associés.

[28]L'Analyse Sémantique Explicite, ou ESA (Egozi O. et al, 2011), est une méthode récemment proposée pour la représentation sémantique de textes en langage naturel de domaine général.

(Egozi O. et al., 2011) ont montré que la représentation *ESA* des concepts ou des requêtes est ambigüe et nécessite un ajustement avant de pouvoir être utilisée en RI. Afin de résoudre ce problème d'ambiguïté, les auteurs ont défini une méthode de Sélection des Fonctionnalités (SF) des concepts candidats.

- **Méthode de *Sélection des Fonctionnalités* (SF) des concepts basée sur la méthode *ESA***

La méthode de *Sélection des Fonctionnalités* (SF) des concepts est basée sur le principe de retour de pertinence utilisée dans la RI où l'utilisateur fournit des jugements de pertinence sur un premier ensemble de concepts récupérés pendant l'étape d'indexation. Ces jugements sont utilisés pour reformuler les requêtes.

La SF est appliquée sur les concepts de la requête pour optimiser leur représentation et supprimer le bruit et l'ambiguïté. Les auteurs ont choisi d'appliquer SF seulement pour les requêtes et non sur les documents car le texte des requêtes est plus court que celui des documents.

Les auteurs ont montré que les résultats obtenus par leur système appelé "*MORAG*" sont meilleurs que les résultats des systèmes existants. Cependant, le principe de retour de pertinence est basé essentiellement sur le jugement des utilisateurs (manuellement) ; d'où une lourdeur cognitive et un risque d'erreur.

2.2.6 Travaux de (Boubekeur F. & Azzoug W., 2013)

(Boubekeur F. & Azzoug W., 2013) proposent une nouvelle approche d'indexation à base de concepts en utilisant *WordNet* et *WordNetDomains*[29]. Cette approche extrait les concepts représentatifs des documents et des requêtes et leur assigne un poids sémantique qui reflète leur importance respective. Le but est de récupérer les documents qui sont sémantiquement pertinents pour une requête utilisateur.

Le processus d'indexation proposé est composé de deux étapes : 1) L'identification des concepts, et 2) La pondération des concepts.

- **Identification des concepts**

Cette étape représente le document (ou la requête) par un index sémantique composé de deux types de termes, des concepts et des mots-clés orphelins :

- Les concepts sont les entrées de *WordNet* (*synsets*) identifiées dans le texte du document.
- Les mots-clés orphelins sont des mots simples (et non vides) du document qui n'ont pas d'entrée dans *WordNet*.

[29] WordNetDomains est une extension du base lexicale Wordnet résultant de l'annotation de chaque synset avec une ou plusieurs étiquettes de domaine d'un ensemble de 176 domaines dans Wordnet.

L'identification des concepts implique deux étapes : a) Identifier les deux types de termes d'indexation du texte, puis les faire correspondre à des *synsets* de *WordNet*, et b) Lever l'ambiguïté des termes ambigus (un terme ambigu correspond à plus d'un *synset*).

Pour lever l'ambiguïté des termes, les auteurs proposent une approche basée sur trois niveaux de désambiguïsation :

- Le premier vise à déterminer la catégorie grammaticale de chaque mot dans un document.
- Le deuxième niveau vise à identifier l'utilisation du domaine d'un mot dans le contexte d'un document (ou de la requête). L'identification de domaine s'appuie sur l'utilisation de *WordNetDomains*. Ce niveau de désambiguïsation limite le nombre de sens traités dans le troisième niveau.
- Le troisième niveau est la désambiguïsation lexicale. Il vise à sélectionner, parmi les sens possibles d'un mot dans le domaine, le concept le plus approprié par rapport au contexte du mot dans le document.

- **Pondération des concepts**

Cette étape attribue un poids à chaque concept identifié (*synset*) qui exprime son importance dans le document. Partant de l'idée que *plus un concept est central au niveau local (dans le document) et central au niveau global (dans la collection) plus il est représentatif du contenu de ce document*, les auteurs de (Boubekeur F. & AzzougW., 2013) ont pondéré les concepts en fonction de leur centralité locale et globale.

La *centralité locale* d'un concept est basée d'une part, sur son importance apparente (mesurée à travers sa fréquence) dans le document et d'autre part, sur son importance discrète dans le document (mesurée à travers sa sémantique liée à d'autres concepts).

La *centralité globale* d'un concept définit son pouvoir de discrimination dans la collection de documents (c'est sa capacité à distinguer les documents qui contiennent des concepts informatifs de ceux qui contiennent des concepts non informatifs).

Les auteurs ont développé un système basé sur la collection de test *TIME*[30]. Les résultats expérimentaux montrent l'efficacité de leur propos par rapport aux approches traditionnelles de RI. Cependant, leur approche présente deux inconvénients :

- L'utilisation d'une base lexicale générale (*WordNet*) pose un problème de couverture terminologique, c'est-à-dire seuls les termes des documents ou des requêtes les plus communs sont référencés par des concepts.

[30]https://isserver11.princeton.edu/

- L'évaluation n'est pas réalisée sur une grande collection de documents telles que *TREC* ou *INEX,* ce qui ne valide que peu l'importance et l'utilité de la nouvelle pondération proposée.

Afin de limiter les problèmes rencontrés lors de l'utilisation de ressources sémantiques généralisées, il convient de se tourner vers des ressources sémantiques spécialisées. Une ressource sémantique spécialisée est une structuration des termes spécifiques à un domaine particulier qui permet de créer une modélisation des connaissances du domaine. La modélisation des connaissances est spécifique à la tâche pour laquelle la ressource sémantique est construite (CeausuV. et Despres S., 2005). C'est en particulier le cas des travaux que nous présentons dans la section suivante.

2.3 Travaux utilisant les ressources sémantiques spécialisées

Dans ce qui suit, nous étudions différents travaux de l'état de l'art traitant d'indexation sémantique basée sur une ressource sémantique spécialisée.

2.3.1 Travaux de (Abascal R., 2005)

Dans (Abascal R., 2005) est définie une approche qui propose un nouveau modèle de document pour les thèses, fondé sur l'utilisation de nouvelles métadonnées. L'approche demande à l'auteur de la thèse de la décrire avec des métadonnées caractérisant le contenu afin de réaliser des recherches pertinentes.

Cette approche vise à permettre l'accès à l'ensemble des thèses par leur contenu sémantique. L'auteur a amélioré de façon notable l'accès au contenu grâce à l'utilisation de « *tags sémantiques* » rajoutés à la thèse. Ceci consiste donc, dans une première phase, à définir les « métadonnées » (concepts) qui permettraient une description plus fine du contenu des thèses. Ensuite, pour définir leur nouvelle structure, chaque partie de la thèse est analysée afin de connaître son organisation liée à la structure sémantique. Cette structure permet d'extraire les éléments les plus porteurs de sens. Grâce à ces éléments, de nouvelles « *métadonnées* » sont définies, puis des « *balise sémantiques* » correspondant à ces métadonnées sont insérés dans la thèse.

L'approche est fondée sur la modélisation sémantique de documents scientifiques, à savoir les thèses de *l'INSA* de Lyon. Cette modélisation est fondée sur : 1) Les concepts issus du contenu de la thèse, et 2) Les concepts qui ont une forte relation avec ceux de la thèse. Pour les premiers concepts, l'auteur a utilisé l'outil *Nomino*[31] de Traitement Automatique de la Langue Naturelle (TALN) capable d'extraire automatiquement des concepts pertinents d'un corpus. Pour les seconds concepts, il a utilisé une ontologie de domaine qui a été spécialement conçue à partir d'un échantillon de thèses en informatique.

[31] www.ling.uqam.ca/nomino.

Ces propositions ont été implantées pour permettre de mieux structurer sémantiquement les thèses de *l'INSA* de Lyon. Une démarche expérimentale a permis aux auteurs de choisir un outil de *TALN* adéquat pour l'extraction automatique des concepts. Ce prototype aide le doctorant pendant la phase de rédaction de sa thèse à ajouter des balises « *sémantiques* ». Le prototype propose trois types de modalités pour ajouter des balises : 1) Selon le choix propre de l'utilisateur, 2) s'appuyant sur la base de concepts, ou 3) en utilisant le logiciel *Nomino* (Patrick Séguéla M., 2001) pour l'extraction des concepts d'un fragment de texte. Grâce à l'utilisation des balises «*balise sémantiques*», le doctorant est capable de mieux organiser sa thèse en évitant les répétitions de concepts. Le prototype permet de restituer à l'utilisateur plusieurs fragments de(s) thèse(s) pertinente(s) pour effectuer une recherche plus ciblée.

Cependant, l'approche proposée est focalisée sur un type particulier de documents, à savoir les thèses. Les auteurs n'ont pas étendu leurs expérimentations à d'autres types de documents.

2.3.2 Travaux de (Harrathi F. et al., 2007)

(Harrathi F. et al., 2007) proposent une méthode d'indexation sémantique automatique pour les documents multilingues. Cette méthode a pour but d'identifier automatiquement les concepts les plus pertinents pour le contenu d'un document ; elle est composée de deux étapes :

- Extraction des termes les plus importants du document en se basant sur les caractéristiques des langues et des méthodes statistiques.
- Extraction des concepts les plus appropriés représentant le contenu du document en utilisant une ontologie multilingue.

La première étape est basée sur l'utilisation du concept *d'information mutuelle*[32], le degré d'association et la fréquence de distribution des termes. Elle se base ainsi sur les propriétés générales de la langue naturelle.

La deuxième étape se décline en quatre sous étapes (cf. Figure 11) : 1) Prétraitement du corpus, 2) Extraction des termes simples, 3) Extraction des termes composées, et 4) Extraction des concepts.

[32]Information mutuelle : mesure la puissance de l'association entre deux mots M1 et M2.

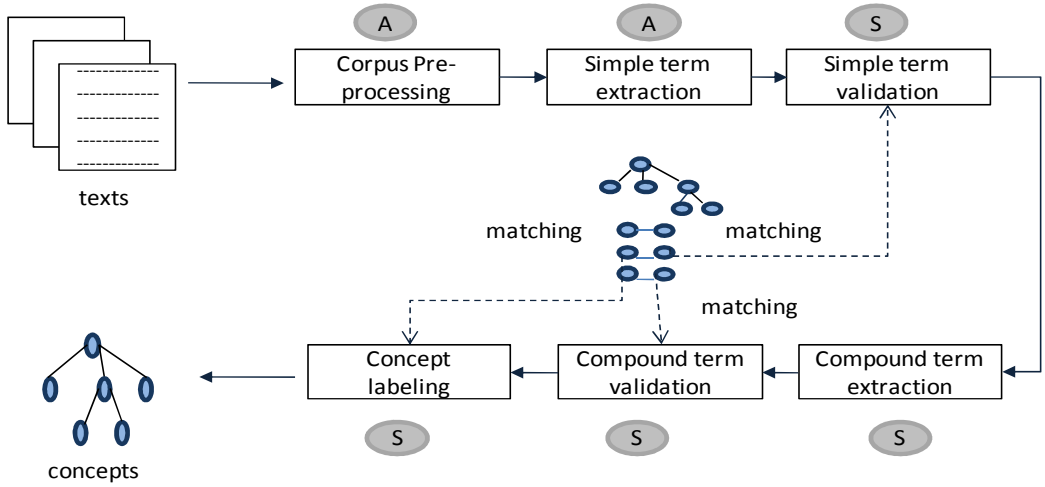


Figure 11. Etapes d'extraction des concepts à partir de textes (Harrathi F. et al., 2007).

Tout d'abord, un prétraitement du corpus restitue une liste de mots non vides avec, pour chaque mot, sa fréquence, sa longueur ainsi que sa position dans le document. Ensuite, une deuxième chaîne de traitements consiste à extraire les termes simples et à les valider. Cette validation est nécessaire car certains termes extraits n'ont pas de correspondance avec l'ontologie. Cela revient à rechercher dans une ontologie les concepts candidats, c'est dire ceux susceptibles d'être associés aux termes extraits. Ces associations sont validées semi-automatiquement par un linguiste. Des termes composés sont ensuite extraits à partir des termes simples validés. Cette extraction est validée selon le même principe que pour les termes simples. Enfin, la dernière sous étape consiste à extraire les concepts candidats à partir du texte du document. Cette étape commence par le classement des termes en des classes en fonction de leurs contextes de distribution dans le corpus. Ainsi, le concept approprié sera recherché dans chaque classe de termes par projection sur la ressource sémantique.

(Harrathi F. et al., 2007) évaluent leur proposition sur le domaine médical à travers la collection *CLEF 2007*. Ils ont testé la performance de la méthode proposée en utilisant des métriques empruntées au domaine de la RI. Durant les expérimentations, ils ont comparé les résultats obtenus par leur méthode et les résultats obtenus avec l'approche linguistique. Ils ont constaté que leur méthode obtient la même performance que l'approche linguistique. De plus, elle présente l'avantage d'être indépendante de la langue des documents. De ce fait, elle est facilement applicable avec d'autres corpus multilingues. Toutefois, la méthode présente deux limites :

- Elle n'est performante que sur des corpus volumineux. En effet, elle est basée sur des mesures statistiques, or ces mesures ne sont significatives que sur des corpus de grandes tailles.
- Un autre inconvénient réside dans l'extraction des concepts candidats à partir du texte du document. Cette correspondance entre termes et concepts consiste à projeter le terme sur la ressource sémantique. Cette projection est stricte et ne prend pas en

considération les variations lexicales et syntaxiques des termes. De plus, cette approche n'utilise pas d'outil d'extraction automatique des concepts.

2.3.3 Travaux de (Dinh D. &Tamine L., 2010)

L'objectif des travaux de (Dinh D. & Tamine L., 2010) est la proposition d'un modèle d'indexation sémantique adapté aux *Dossiers Médicaux des Patients* (*DMP*) en utilisant le thésaurus *MeSH*. Ce modèle servira de support à des processus de recherche d'information médicale, permettant d'encourager l'expérience collective des médecins. Plus précisément, la contribution porte sur l'indexation sémantique de l'information explicite (et non implicite) contenue dans le texte des documents qui composent le *DMP* et ce, en proposant :

- Une méthode pour la désambiguïsation des descripteurs sémantiques issus de *MeSH*, en tenant compte de leur contexte local dans le document. Comparativement aux autres méthodes de désambiguïsation dans le domaine biomédical, cette méthode est basée sur le contexte local des concepts dans le document en exploitant la hiérarchie de *MeSH* pour identifier le sens correct et non les métadonnées.
- Un index sémantique construit selon un schéma de pondération qui combine la spécificité des concepts dans les documents et leur centralité dans les dossiers.

Cette méthode est composée des deux étapes de la Figure 12 : 1) *Annotation sémantique*, suivie de 2) *Génération de l'index sémantique*.

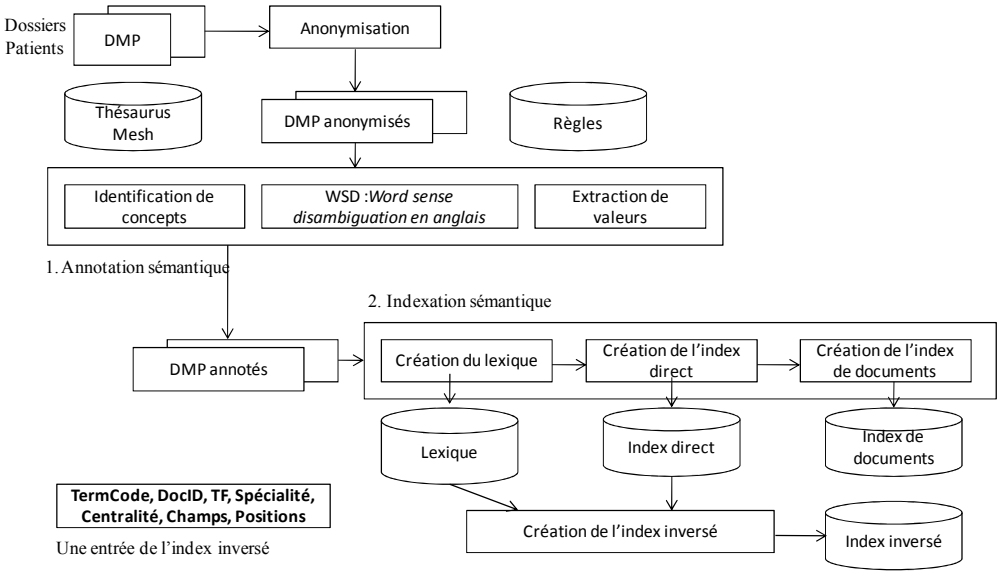


Figure 12. Processus d'indexation sémantique de dossiers médicaux (Dinh D. & Tamine L., 2010).

L'étape d'annotation sémantique est composée de deux objectifs : 1) Identifier les concepts associés aux termes extraits de *MeSH* ainsi que les valeurs associées dans le document, 2) désambiguïser les concepts.

L'identification puis l'annotation conceptuelle des documents sont basées sur l'association des termes de longueur maximale aux entrées de *MeSH*. Cette annotation est suivie d'une étape de désambiguïsation afin de résoudre le problème de l'ambiguïté liée à la polysémie des termes. La désambiguïsation est fondée sur :

- L'unicité du sens d'un terme qui désigne un concept spécifique dans le document (Gale W. A. &al., 1992).
- La corrélation des sens des termes voisins : Les sens associés à des termes voisins (sur une fenêtre) sont sémantiquement proches les uns des autres.
- La priorité du sens est définie par la position des termes désignant les concepts : Le concept qui se trouve le plus à gauche détermine le sens global de la suite du discours. Ce principe est inspiré par la notion de chaîne sémantique du discours (Morris J. et Hirst G., 1991). Cette dernière a été définie comme une séquence de mots $w_1, w_2, ..., w_i, w_{i+1} ..., w_n$ reliés sémantiquement dans un texte de sorte à ce que le mot w_i soit relié au mot w_{i+1} par une relation lexico-sémantique.

La deuxième étape, nommée génération de l'index sémantique, a pour objectif de générer un index sémantique contenant à la fois les concepts identifiés selon l'approche de désambiguïsation précédente et les termes qui ne correspondent pas à des entrées de *MeSH*.

Dans le but de mettre en évidence l'importance des concepts dans les dossiers médicaux, les auteurs de la méthode ont proposé un descripteur défini par un schéma de pondération spécifique. Ce schéma tient compte du niveau de description de l'index ainsi que de la localisation des concepts dans le document et dans la hiérarchie de *MeSH*, ceci dans le but de traduire à la fois leur spécificité et leur centralité.

Pour évaluer ce modèle d'indexation, deux séries d'expérimentations ont été menées ; la première série utilise une indexation classique basée sur les termes simples alors que la seconde série d'expérimentation est basée sur l'application d'une indexation sémantique aux documents et aux requêtes. Les résultats obtenus montrent une amélioration (+15.67%) de l'approche d'indexation sémantique par rapport à l'approche d'indexation classique qui ne prend en considération que des mots séparés. Ceci montre clairement l'intérêt de l'indexation sémantique et celui du schéma de pondération proposé.

Toutefois, une partie importante de l'évaluation est effectuée manuellement. D'où une lourdeur cognitive et un risque d'erreur car la tâche de l'utilisateur est importante.

2.3.4 Travail de (Bevan K. et al., 2012)

Ce travail expose une méthode de construction de graphe pondéré pour les concepts extraits en utilisant l'ontologie du domaine médical *SNOMED CT*[33] pour "*Systematized*

[33]www.nlm.nih.gov/snomed/

Nomenclature of Medicine--Clinical Terms". Dans ce graphe, les nœuds représentent les concepts et les arcs entre les nœuds représentant les relations.

Dans la littérature, il existe deux types de travaux traitant de représentation en graphe de pondération : 1) Graphe de pondération de termes extraits d'un document et 2) Graphe de pondération de concepts extraits d'un document en utilisant une ressource sémantique.

Les approches construisant des graphes de pondération de termes représentent chaque document par un graphe où les sommets sont les termes et les arcs sont les relations entre les termes. Ces relations peuvent être de cooccurrence simple dans un contexte ou bien des relations grammaticales plus complexes. L'importance d'un terme dans un document peut alors être estimée par le nombre de termes du document et leur importance respective. Cette représentation en graphe ne capte pas les dépendances entre termes ce qui constitue un inconvénient majeur lors de l'interprétation de textes médicaux.

Ce travail de (Bevan K. et al., 2012) est inspiré des approches de pondération des documents en utilisant des concepts médicaux. Cependant, les auteurs ont constaté que lorsqu'ils adoptent le même principe de construction, la pondération ne tient pas compte de l'importance du concept dans le domaine médical puisqu'elle est calculée uniquement en fonction du document.

Pour pallier ce problème, les auteurs ont intégré la connaissance contenue dans l'ontologie *SNOMED CT* dans la fonction de pondération des concepts du graphe des concepts. Cette intégration possède un double avantage : premièrement, elle améliore la RI en augmentant la restitution des documents pertinents ; deuxièmement, elle permet l'identification des concepts importants dans le domaine médical de référence. Par exemple, dans la Figure 13, le concept "*Asthma*" (asthme en français) est relié à 50 autres concepts de l'ontologie *SNOMED CT* ; ce qui montre que l'indication d'importance de ce concept est élevée.

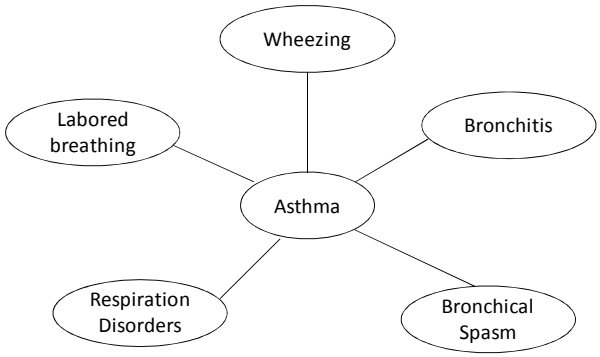


Figure 13. Relations du concept "*Asthma*" avec d'autres concepts (Bevan K. et al., 2012).

Les auteurs ont intégré cette indication d'importance dans le nouveau graphe de pondération des concepts. Cette méthode prend en compte le nombre de concepts reliés dans toute l'ontologie médicale *SNOMEDCT* et non pas seulement dans le document. Ainsi, l'importance du concept est devenue plus pertinente car elle est calculée par rapport au

domaine médical de référence sans se limiter au corpus lui-même ; ce qui est positif pour l'utilisateur.

L'avantage majeur de cette approche est qu'elle permet d'exclure un grand nombre de termes d'une requête utilisateur, ce qui réduit la forme de la requête et conduit à une amélioration de la pertinence du résultat fourni.

L'inconvénient de cette approche réside dans la pondération des concepts qui ne prend pas en compte la structure hiérarchique des concepts dans la ressource sémantique utilisée. De fait, la sémantique de la relation hiérarchique n'est pas utilisée.

2.3.5 Travaux de (Majdoubi J. et al., 2012)

Dans (Majdoubi J. et al., 2012), les auteurs proposent également une approche d'indexation conceptuelle d'articles médicaux en utilisant le thésaurus *MeSH*.

Cette approche est schématisée dans la Figure 14 et se compose de quatre étapes principales : a) Prétraitement, b) Extraction de termes, c) Pondération des termes, et d) Sélection des descripteurs.

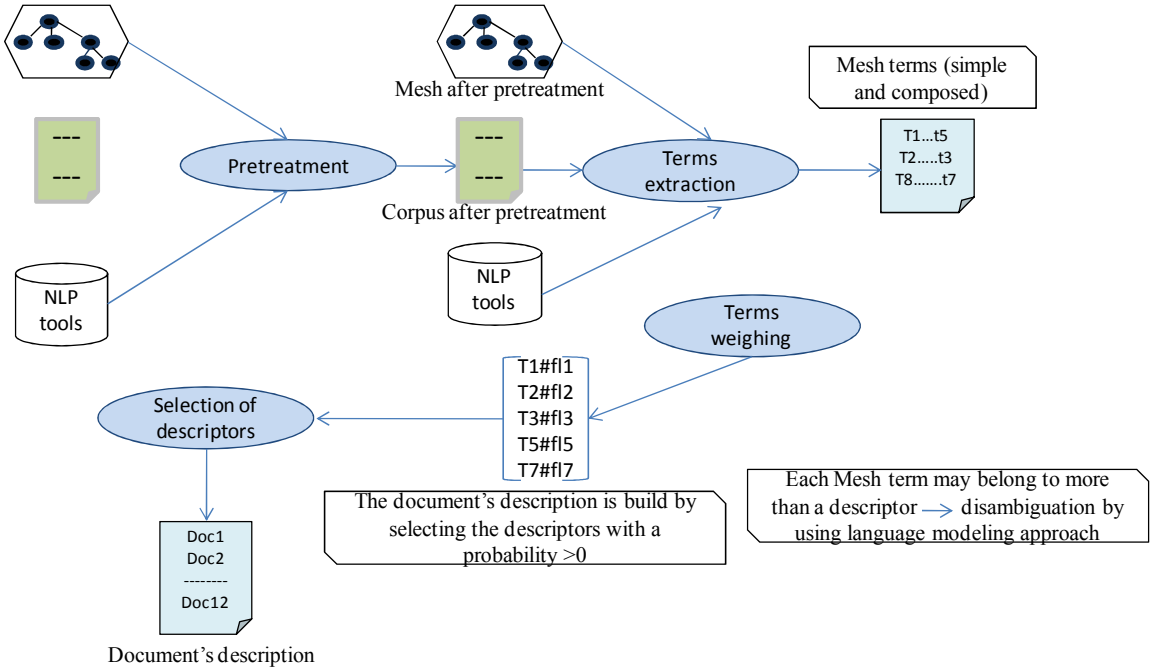


Figure 14. Architecture de l'approche d'indexation (Majdoubi J. et al., 2012).

La première étape de prétraitement consiste à découper le texte en un ensemble de mots. Les auteurs utilisent l'outil de segmentation *TreeTagger* (Cunningham H. et al., 2002). Ensuite, *TreeTagger* permet d'identifier la catégorie grammaticale (nom, verbe, ...) et le lemme (radical) de chaque mot extrait. Enfin, l'étape d'élimination des mots vides est réalisée pour chaque article médical du corpus.

La deuxième étape permet l'extraction de termes en utilisant trois catégories de techniques : linguistique, statistique, et hybride (linguistique et statistique).

La troisième étape correspond au calcul des poids des termes selon une nouvelle technique de pondération. Pour déterminer l'importance des termes, cette technique exploite les relations dans *MeSH*, plutôt que de compter uniquement leur fréquence statistique. Plus précisément, un poids supplémentaire est attribué aux termes qui sont extraits du titre ou du résumé du document. Pour ce faire, deux mesures sont utilisées :

- Le *Poids du Contenu de la Structure* (CSW pour *Content Structure Weight*) : cette mesure tient compte de la fréquence du concept dans chaque partie du document, plutôt que dans l'ensemble du contenu document. La mesure d'un concept tient compte de l'occurrence de ce concept, ainsi que de son emplacement dans le document. Par exemple, un concept du « titre » aura un coefficient plus élevé (coefficient de 10 par exemple) qu'un concept figurant dans le « résumé » (coefficient de 8 par exemple).
- Le *Poids Sémantique* (SW pour *Semantic Weight*) : Le poids sémantique d'un terme t_i dans un document *d* dépend de ses synonymes dans l'ensemble des termes candidats générés par l'étape d'extraction des termes. Enfin, la sélection des descripteurs liés à un terme donné est réalisée à l'aide d'un *modèle de langue* [34] (Sparck-Jones K. et al, 2003) (Boughanem M. et al., 2004) (Fuhr N., 2000). Tout d'abord, (Majdoubi J. et al., 2012) ont estimé un modèle de langue pour chaque document de la collection et pour chaque descripteur de *MeSH*. Ensuite, ils ont classé les documents par rapport à la probabilité que le modèle de langue d'un document génère le descripteur *MeSH*. Le principe de base de sélection de descripteurs est « *pour un terme t_i donné ayant plusieurs sens dans le document d, le meilleur descripteur est celui qui a la plus forte probabilité d'être généré par le modèle de langue de ce document* ».

Cette approche d'indexation a été expérimentée sur un ensemble de 500 articles médicaux. Une évaluation expérimentale et une comparaison de leur système "*BIOINSY*" avec d'autres outils d'indexation confirment l'avantage d'utiliser le modèle de langue dans le processus d'indexation conceptuelle.

Un inconvénient de cette approche réside toutefois dans l'extraction des concepts candidats qui reste limitée à explorer uniquement le contenu textuel du document, sans recours à un référentiel sémantique. La mise en correspondance entre terme et concept consiste à projeter le terme sur la ressource sémantique. Il s'agit d'une projection stricte qui ignore les variations lexicales et syntaxiques des termes.

[34]Dans le modèle de langage, la requête est inférée par l'utilisateur à partir de ces documents. Un document n'est pertinent que si la requête utilisateur ressemble à celle inférée par le document. Par conséquence, ce modèle cherche à estimer la probabilité que la requête soit inférée (générée) par le modèle du document.

2.4 Bilan et synthèse

Nous présentons, dans le tableau 3, une synthèse des travaux relatifs à l'indexation sémantique de documents étudiés dans ce chapitre. Les colonnes représentent les critères d'évaluation et les lignes sont les travaux étudiés.

Approche d'indexation sémantique	Automatique	Type des documents	Ressource sémantique	Modèle de structuration sémantique
(Zargayouna H. et al., 2004)	Oui	Document XML structuré	Générale *WordNet*	Réseau sémantique
(Kang B. Y. & Lee S., 2005)	Oui	Non Structurés	Générale *WordNet*	Réseau sémantique
(Baziz M. et al., 2007)	Oui	Non Structurés	Générale *WordNet*	Réseau sémantique
(Tagarelli A. & Grec S., 2010)	Oui	Document XML structuré	Générale *WordNet*	Structure sémantique
(Egozio et al, 2011)	Oui	Non Structurés	L'analyse Sémantique Explicite : domaine général	Analyse Sémantique Explicite
(Boubekeur F. & Azzoug w., 2013)	Oui	Non Structurés	Générale *WordNet*	Graphe conceptuel
(Abascal R., 2005)	Semi	Non Structurés	Spécialisée Ontologie de domaine	Structure sémantique
(Harrathi F. et al., 2007)	Semi	Document XML structuré	Spécialisée *MeSH*	Graphe conceptuel
(Dinh D. & Tamine L., 2010)	Oui	Non Structurés	Spécialisée *MeSH*	Graphe conceptuel
(Bevan k. et al., 2012)	Oui	Non Structurés	Spécialisée *SNOMED CT*	Graphe conceptuel
(Majdoubi J. et al., 2012)	Oui	Non Structurés	Spécialisée *MeSH*	Graphe conceptuel

Tableau 3. Tableau comparatif des travaux traitant d'indexation sémantique.

Nous remarquons que certaines des approches précédemment étudiées supposent que l'utilisateur est expert dans le domaine (critère 1) et que l'ensemble des documents partagent une même ressource sémantique (Critère 4). Notons aussi que plusieurs travaux ont utilisé *WordNet* qui représente une base lexicale très générale et non formalisée pour modéliser un domaine donné, ce qui peut poser des problèmes de couverture. D'autres approches se sont intéressées uniquement à l'enrichissement des noms de balises XML sans se baser sur le contenu textuel des documents comme ceux de (Tagarelli A. & Grec S., 2010). Certains travaux, comme ceux de (Harrathi F. et al., 2007), (Baziz M. et al., 2007) et (Zargayouna H.

et al., 2004), se sont intéressés au domaine de la RI ; ils sont arrivés à améliorer le nombre de documents pertinents restitués par rapport à une requête utilisateur, en utilisant une ontologie.

En complément des travaux abordant l'aspect sémantique de contenus textuels, nous proposons une *modélisation sémantique* des documents XML. Plus précisément, nous proposons une approche de détermination d'une structure sémantique par document XML en se basant sur sa structure logique et son contenu. La structure sémantique d'un document organise un ensemble de concepts traduisant le contenu sémantique des éléments de la structure logique. La spécificité de nos travaux réside dans le fait que les concepts sont pondérés par rapport à leur organisation dans la structure hiérarchique de la ressource sémantique ; c'est-à-dire que le niveau hiérarchique d'un concept dans la ressource sémantique est important : plus on descend en profondeur dans la hiérarchie, plus le concept est important et par conséquent son poids sera élevé. En effet, les concepts sont pondérés de manière à donner plus d'importance aux concepts les plus spécifiques (i.e., localisés en bas de la hiérarchie) car ces éléments présentent généralement une information plus fine (ciblée) et plus spécifique. Nos travaux se différencient ainsi de ceux de la littérature, comme (Baziz M. et al. 2007) (Harrathi F. et al., 2007) (Bevan k. et al., 2012) (Boubekeur F.& Azzoug w., 2013) (Zargayouna H. et al., 2004), qui ne prennent pas en compte cette structure hiérarchique des concepts dans la ressource sémantique utilisée.

Il est important de noter que nous avons choisi de représenter la sémantique des documents par une structure arborescente basée sur la structure logique plutôt que sur un réseau sémantique ou un graphe conceptuel en raison des inconvénients de ces modèles. Les réseaux sémantiques sont exprimés par une représentation simple, mais ils sont rarement appliqués à des problèmes pratiques en raison de leur difficulté d'utilisation. Le modèle des graphes conceptuels soulève quant à lui une autre difficulté inhérente à la définition des nœuds et des liens. Dans la structure sémantique, les données sont organisées en fonction de leur sens et de leur définition. L'identification d'une structure sémantique simple permet donc un accès plus facile à l'information.

2.5 Conclusion

Nous avons étudié dans ce chapitre un ensemble significatif de travaux traitant de l'indexation sémantique. Nous avons classé ces travaux selon le type de ressource sémantique utilisée : *généralisée* ou *spécialisée*.

Nous avons noté que dans la majorité des travaux traitant d'indexation sémantique, la pondération des concepts utilise des mesures statistiques. Ces mesures exploitent des informations sur les descripteurs de ces concepts, ainsi que sur leur répartition dans le document et dans le corpus. Cependant, la plupart des travaux n'accorde pas d'importance, lors de la pondération des concepts d'une ressource sémantique, à la position de ces concepts

dans la hiérarchie, et ainsi ne différencient pas le fait qu'un concept soit plus générique (peu profond) ou plus spécifique (assez profond). Afin d'introduire cette notion, nous proposons de pondérer les concepts d'une ressource sémantique de manière à donner plus d'importance aux concepts les plus spécifiques.

Afin de renforcer la prise en compte des aspects sémantiques du contenu, nous proposons dans le chapitre suivant une approche originale visant à déterminer la structure sémantique d'un document *XML*. Cette approche sera basée conjointement sur la structure logique du document, sur son contenu et sur une ressource sémantique spécialisée. Plus précisément, l'approche part de la structure logique d'un document, puis la transforme en remplaçant chaque nœud par le concept le plus approprié, c'est-à-dire celui qui correspond aux termes significatifs extraits du texte associé à ce nœud.

Chapitre 3

Approche de détermination d'une structure sémantique par document XML

Sommaire

3.1 Introduction

La nature des sources d'information évolue, et les documents numériques traditionnels plats ne contenant que du texte s'enrichissent d'informations structurelles et multimédia. Cette évolution est accélérée par l'expansion du Web, et les documents semi-structurés de type XML représentent une grande partie des documents numériques mis à disposition des utilisateurs. Des travaux assez récents soulignent l'importance du contenu textuel de ces documents pour la prise de décisions (Sauvagnat K., 2005). Cependant, les balises de ces documents sont souvent sémantiquement pauvres dans le sens où elles ne sont pas informationnelles ; de plus, le découpage concerne généralement la structure logique du document (*Titre*, *Sections*, *Paragraphes*) et n'est que très rarement lié à la sémantique du contenu. Il en résulte que l'exploitation sémantique de ces documents est quasiment hors de portée des décideurs voire même des informaticiens. Le besoin d'une structuration sémantique des documents se fait de plus en plus sentir ; dès lors, la communauté scientifique oriente certains de ses travaux de recherche vers cet objectif. C'est également dans ce contexte que nous nous plaçons. Plus particulièrement, l'objectif de ce chapitre est de proposer une approche pour la détermination d'une structure sémantique par document XML. Notre idée de base est d'identifier les concepts incarnés dans un document XML et de les structurer sémantiquement. Cette structuration sémantique ne peut se faire sans le recours à un référentiel de base du domaine thématique des documents qui décrit d'une façon plus ou moins standard des liens, au moins hiérarchiques, entre concepts. Ce référentiel pourrait être toute forme de ressources sémantiques telles qu'une ontologie de domaine, une taxonomie… Dans notre cas, nous avons opté pour l'utilisation d'une taxonomie dans la mesure où ces types de ressources sont plus faciles à construire et à trouver.

Dans ce chapitre, nous présentons notre approche de détermination d'une structure sémantique d'un document XML centré-texte. Nous commençons par décrire le contexte de travail, le processus de construction de la structure sémantique, puis nous détaillons les différentes étapes de l'approche proposée.

3.2 Présentation générale de l'approche proposée

Dans (Khrouf et al., 2011) et (Ben Messaoud et al., 2011), les auteurs ont proposé une approche pour la classification et l'analyse multidimensionnelle de documents. Ces travaux regroupent les documents XML en fonction de leur structure logique. Les structures logiques identiques ou similaires de documents sont alors regroupées et décrites par des structures génériques. L'approche proposée a été vérifiée et validée pour les documents XML orientés-données (contenant généralement peu de texte). Nous souhaitons étendre ces travaux pour les documents orientés-textes (rapports, articles scientifiques, news…). A cette fin, nous

proposons de dériver, à partir de la structure logique et du contenu d'un document XML orienté-texte, une structure qui reflète sa sémantique. L'objectif principal est d'obtenir une structure sémantique pour document XML en s'appuyant sur sa structure spécifique et en se référant à une ressource sémantique spécialisée qui pourrait être une taxonomie de domaine.

La Figure 15 schématise les différents niveaux de représentation d'un document selon notre approche. La Figure 15.1 est un exemple de document orienté-texte bien formé en XML appelé *XML-paper*. La Figure 15-a décrit la structure logique spécifique *Sp1-paper* et la Figure 15-b décrit la structure sémantique *SEM1-paper* associée à la structure logique *Sp1-paper* du document XML *XML-paper*.

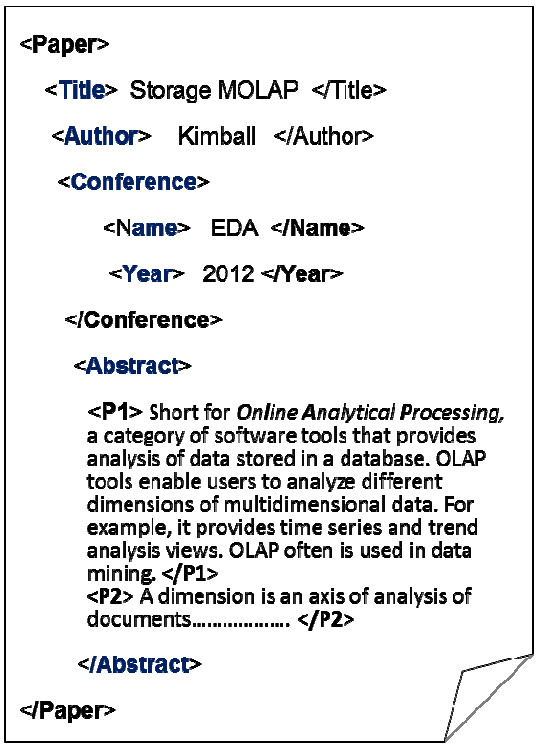


Figure 15-a : XML-paper

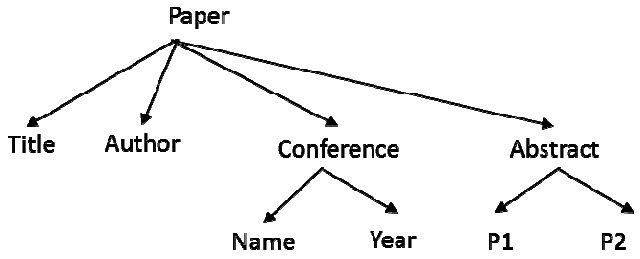


Figure 15-b : Sp-paper
(structure spécifique de XML-paper)

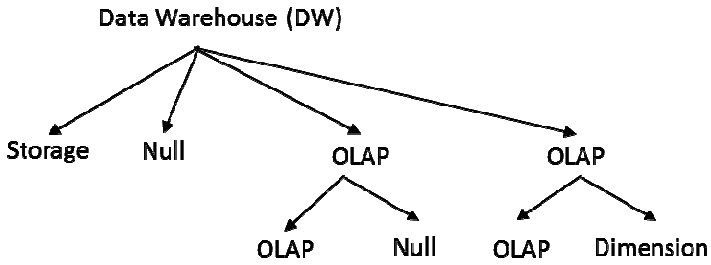


Figure 15-c : -SEM-paper
(structure sémantique de XML-paper)

Figure15. Exemple de document XML (Figure15-a) avec sa structures logique (Figure15-b) et sa structure sémantique (Figure15-c).

Dans cet exemple, la structure spécifique *Sp-paper* est constituée de quatre éléments à savoir : *Title*, *Author*, *Conference* (décrit par *Name* et *Year*) et *Abstract* (ayant deux paragraphes *P1* et *P2*). La structure sémantique *SEM-paper* associée à *XML-paper*est un arbre dérivé de sa structure logique *Sp-paper* et où chaque élément de type contenu textuel est remplacé par un concept issu de la taxonomie ; ainsi par exemple, *Title* est remplacé par *Storage*, et *Conference* par *OLAP*. Les éléments *Author* et *Year*, ont quant à eux été remplacés par le concept Null dans la mesure où ils ont été considérés ici comme des éléments de type métadonnées et non comme des éléments de type contenu textuel.

Seuls les éléments ayant des contenus textuels se verront associer un concept à partir de la ressource sémantique en fonction des mots-clés extraits du contenu textuel de chaque élément. Les éléments considérés comme étant des métadonnées se verront associé à un concept « Null » de sorte à ce que la structure sémantique reste conforme à la structure logique qui a permis de l'engendrer.

La Figure 16 illustre l'approche que nous proposons pour la détermination de la structure sémantique d'un document.

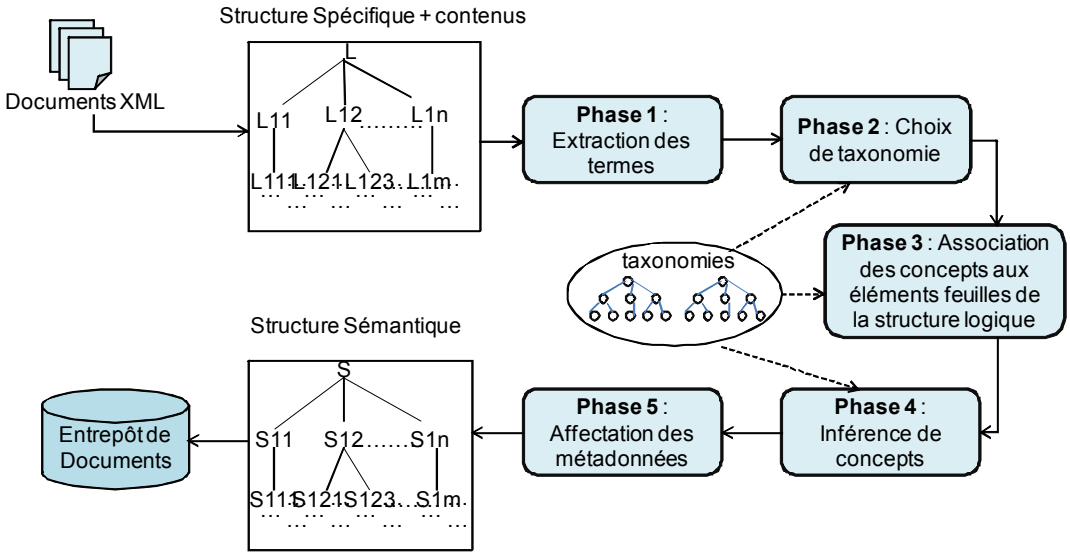


Figure 16. Approche de détermination d'une structure sémantique (Ben Meftah S. et al., 2012)

Cette démarche se décline en cinq phases :

- ***Phase 1*** : Extraction des termes. Il s'agit d'extraire les mots-clés significatifs des éléments feuilles du document (fragments textuels associés aux éléments feuilles). L'extraction des mots-clés repose sur un processus d'indexation classique, tel que défini en recherche d'information.
- ***Phase 2*** : Choix d'une taxonomie. L'objectif ici est de déterminer, parmi un ensemble de taxonomies de domaines, celle qui correspond au mieux au domaine du document et qui sera par conséquent la taxonomie de référence pour construire la structure sémantique du document. Le choix d'une taxonomie est basé sur les mots-clés extraits lors de la phase 1.
- ***Phase 3*** : Association d'un concept à chaque élément feuille de la structure spécifique du document. Il s'agit de chercher, dans la taxonomie de domaine retenue en phase 2, le concept le "plus approprié" pour la description de la sémantique de l'élément feuille

et ceci en tenant compte des mots-clés qui le décrivent. Le concept trouvé est alors assigné à l'élément feuille.

- ***Phase 4*** : Inférence des concepts. La construction de la structure sémantique pour un document se base sur la structure spécifique du même document. Or dans la phase précédente seuls les éléments-feuilles textuels (paragraphes ou sections) se sont vu assigner des concepts ; c'est-à-dire que les nœuds non-feuilles n'ont pas encore de concept associés. Dans notre démarche, les concepts des éléments feuilles seront ré-exploités pour inférer les concepts à associer à leur ascendant. Cette phase utilise la taxonomie sélectionnée et est réitérée pour tous les niveaux de la structure sémantique du document jusqu'à atteindre la racine.
- ***Phase 5*** : Affectation des métadonnées. Elle consiste à reprendre les balises de la structure spécifique qui représentent des métadonnées (i.e. *Author*, *Year*). Les métadonnées utilisées lors de cette phase sont celles du Dublin Core. Nous avons choisi de réaliser cette phase à la fin de la démarche proposée et non au début afin de permettre de remplacer certaines métadonnées (i.e. *Title*, *Abstract*) par les concepts représentant la sémantique de leur contenu textuel.

Les phases 2, 3, 4 et 5 sont détaillées dans les sections suivantes. Pour la réalisation de la première phase, nous recourons aux techniques de recherche d'information (Salton G. & McGill M.J., 1983) (Soulé-Dupuy C., 2001).

3.3 Choix d'une taxonomie pour un document

Dans notre approche, un entrepôt de documents peut contenir des documents appartenant à plusieurs domaines et dispose d'un ensemble de taxonomies spécifiques à ces domaines. Afin d'apporter de la sémantique aux différents éléments constituant un document, nous recourons à ces taxonomies de l'entrepôt.

Dans nos travaux, nous avons opté pour l'utilisation de taxonomies (cf. Chapitre 1) en tant que ressources terminologiques spécifiques à un domaine. A partir d'une taxonomie, nous identifions un ensemble de concepts et nous utilisons les relations hiérarchiques entre ces concepts pour l'inférence de concepts pour les nœuds non feuilles de la structure du document.

3.3.1 Pondération des taxonomies et de leurs concepts

Rappelons que notre objectif est de créer une structure sémantique, essentiellement en se basant sur une taxonomie de domaine choisie parmi plusieurs. Se pose alors le problème du choix d'une taxonomie. Il s'agit de répondre à la question : *Quelle est la taxonomie la mieux appropriée pour déterminer les concepts de la structure sémantique d'un document* ?

Pour cela, nous ne devons pas considérer comme équi-importantes les différentes taxonomies. En effet, étant donné un élément d'un document (e.g., Paragraphe) et deux concepts susceptibles d'être associés, situés à deux niveaux hiérarchiques différents dans une taxonomie, nous préférons associer le concept-fils (i.e. le plus spécifique) car il présente une information plus précise que le concept-père. Ainsi, il nous semble significatif de pondérer les concepts des taxonomies de manière à donner plus d'importance aux concepts les plus spécifiques (situés en bas de la hiérarchie,) car ils sont plus précis que leurs ascendants.

Le principe de pondération des taxonomies que nous avons proposé est le suivant :

- Affecter des coefficients aux concepts de la taxonomie par niveaux par niveaux (de sorte à ce que les concepts de plus bas niveau aient des coefficients plus importants que les concepts de niveau inférieur).
- Calculer une marge ε qui constitue la différence entre le poids du concept du niveau $i+1$ et le poids du concept du niveau i.
- Calculer le poids de base λ de chaque concept de la taxonomie.
- Calculer le poids effectif de chaque concept. Il est égal à :

Poids de base + (Marge * Coefficient)

Pour effectuer cette pondération, nous sommes confrontés au problème de la taille des taxonomies ou plus précisément de la différence entre les tailles des taxonomies à comparer. En effet, nous pouvons rencontrer des taxonomies plus détaillées et plus élaborées que d'autres (cf. Figure 17). Évidemment, ces taxonomies ne doivent pas avoir le même poids, nous tenons compte de la taille de chaque taxonomie lors de la répartition de son poids sur ses différents concepts. Ainsi, nous évitons qu'un concept appartenant à une taxonomie de petite taille reçoive un poids nettement supérieur à un autre concept se trouvant dans une taxonomie de grande taille.

Par exemple, si chacune des deux taxonomies O_1 et O_2, possédant respectivement 4 et 10 concepts, a un poids égal à 1, alors chaque concept de O_1 aura un poids de 1/4 alors que le poids de chaque concept de O_2 sera de 1/10.

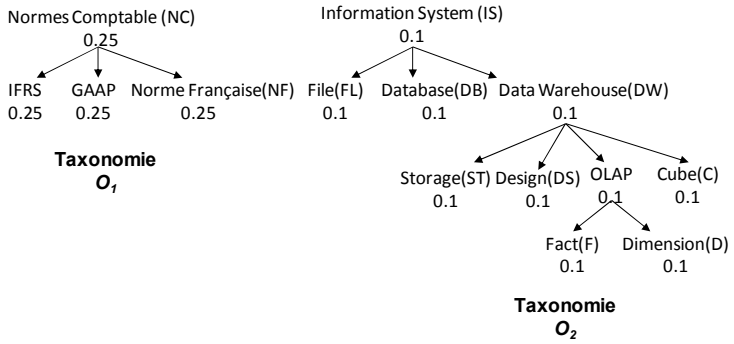


Figure 17. Pondération non discriminante des concepts des taxonomies O_1 et O_2.

Cette pondération non discriminante est à éviter car les concepts des taxonomies moins élaborées (i.e., contenant peu de concepts) auront plus d'importance et seront par conséquent privilégiés. Pour éviter ce problème, nous calculons le poids de chaque taxonomie en fonction du nombre de ses concepts et du nombre total de concepts dans toutes les taxonomies qui seront comparées (i.e. utilisées dans la phase de *Choix de taxonomie*). Nous calculons alors le poids, noté PO_k, d'une taxonomie O_k selon la Formule 1 suivante :

$$PO_k = \frac{|O_k|}{\sum_{i=1}^{N} |O_i|} * \qquad [1]$$

Où :

- PO_k est le poids de la taxonomie O_k,
- $|O_i|$ est le nombre de concepts dans la taxonomie O_i, et
- N est le nombre des taxonomies d'un domaine, disponibles dans l'entrepôt de documents.

Reprenons l'exemple de la Figure 17, après pondération selon la Formule 1, nous obtenons les poids suivants : PO_1 = 0,571 et PO_2= 1,429. Ce qui donne plus d'importance à la taxonomie O_2, plus complète qu'O_1.

A ce stade, nous répartissons le poids de chaque taxonomie entre ses différents concepts de manière à donner plus d'importance aux éléments les plus spécifiques dans l'arborescence. Ainsi, pour un concept C situé à un niveau i d'une taxonomie, tous ses sous-concepts immédiats (niveau $i+1$) auront des poids supérieurs à celui de C. L'affectation des coefficients est réalisée par la fonction $Coeff(C_i,O_k)$ suivante :

$$: \quad Coeff(C_i, O_k) \begin{cases} = i & \text{si } C_i \text{ est racine d'un sous arbre, à savoir élément non feuille de niveau } i \\ = l_{max} + (1-2) & \text{si } C_i \text{ est un élément feuille de niveau} \end{cases} \qquad [2]$$

Où :

i_{max} correspond à la profondeur maximale de l'arbre.

Par exemple : i_{max} = 4 pour O_2 *et* i_{max} = *2 pour* O_1.

Exemple (suite). Dans notre exemple courant, cela se traduit par les coefficients suivants :

- Chaque élément non feuille, aussi appelé *père*, reçoit un coefficient égal à son niveau dans la taxonomie. Ainsi, dans la Figure 18, les nœuds "*Information System*", "*Data Warehouse*" et "*OLAP*" sont des nœuds pères, et les coefficients 1, 2 et 3 leur seront respectivement affectés en fonction de leur position (niveau) en partant de la racine (qui se situe au niveau 1).

– Chaque élément feuille reçoit un coefficient, en fonction de son niveau, égal à $i_{max}+(i-2)$. Dans notre exemple (cf. Figure18), nous affectons le coefficient 4 aux concepts "*File*" et "*DataBase*" de niveau 2, le coefficient 5 aux concepts "*Storage*", "*Design*" et "*Cube*" de niveau 3, etc.

$$Coeff(NC, O_1) = i = 1$$
$$Coeff(IFRS, O_1) = i_{max} + (i-2) = 2 - (2-2) = 2$$
$$Coeff(GAAP, O_1) = i_{max} + (i-2) = 2 - (2-2) = 2$$
$$Coeff(NF, O_1) = i_{max} + (i-2) = 2 - (2-2) = 2$$

$$Coeff(IS, O_2) = i = 1$$
$$Coeff(DW, O_2) = i = 2$$
$$Coeff(OLAP, O_2) = i = 3$$
$$Coeff(FL, O_2) = i_{max} + (i-2) = 4 + (2-2) = 4$$
$$Coeff(DB, O_2) = i_{max} + (i-2) = 4 + (2-2) = 4$$
$$Coeff(ST, O_2) = i_{max} + (i-2) = 4 + (3-2) = 5$$
$$Coeff(DS, O_2) = i_{max} + (i-2) = 4 + (3-2) = 5$$
$$Coeff(C, O_2) = i_{max} + (i-2) = 4 + (3-2) = 5$$
$$Coeff(F, O_2) = i_{max} + (i-2) = 4 + (4-2) = 6$$
$$Coeff(D, O_2) = i_{max} + (i-2) = 4 + (4-2) = 6$$

La Figure 18 montre les coefficients ainsi affectés pour les taxonomies O_1 et O_2.

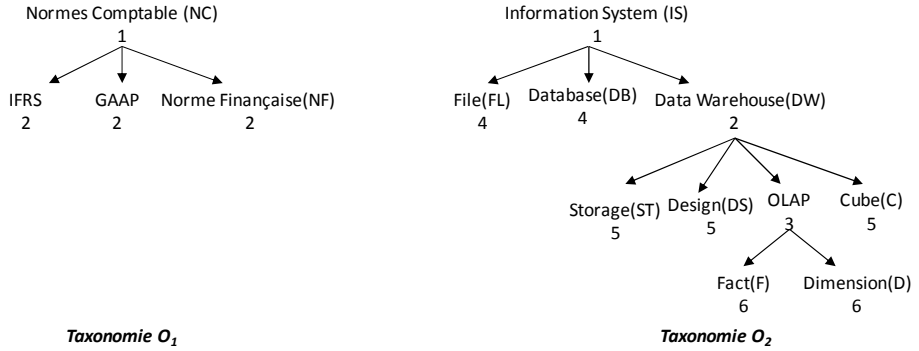


Figure 18. Coefficients des concepts de la taxonomie O_2.

A ce niveau, nous calculons une marge par taxonomie notée ε_k (cf. Formule 3) qui sera utile pour la détermination des poids des concepts de la taxonomie O_k. En fait, ε_k constitue la marge entre un concept appartenant à un niveau $i+1$ de la taxonomie O_k par rapport à un concept d'un niveau i d'O_k dont l'objectif est de favoriser les descendants.

$$\varepsilon_k = \frac{PO_k}{SO(O_K)} \quad avec \quad SO(O_K) = \sum_{|O_k|}^{i=1} Coeff(C_i, O_k) \quad [3]$$

Où :

– PO_k est le poids de la taxonomie O_k calculé selon la Formule 1,

– $SO(O_k)$ est la somme des coefficients de la taxonomie O_k,

– $Coeff(C_i, O_k)$ est le coefficient du concept C_i dans la taxonomie O_k,

- $|O_k|$ est le nombre de concepts dans la taxonomie O_k.

Exemple (suite). Dans le cas des taxonomies O_1 *et* O_2 (Figure 18), la valeur de ε_1 et ε_2 sont les suivantes :

$$\varepsilon_1 = \frac{PO_1}{SO(O_1)^2} = \frac{0.571}{(1+2+2+2)^2} = \frac{0.571}{7^2} = 0.01165$$

$$\varepsilon_2 = \frac{PO_2}{SO(O_2)^2} = \frac{1.429}{(1+4+4+2+5+5+3+5+6+6)^2} = \frac{1.429}{41^2} = 0.0008$$

Nous déterminons à ce niveau le poids de base d'un concept, noté λ_k , c'est-à-dire sans tenir compte des coefficients des concepts. Il s'agit de soustraire du poids de la taxonomie l'ensemble de toutes les marges ($\varepsilon_k * SO(O_k)$), et de diviser par le nombre de concepts.

La formule pour calculer le poids de base λ_k d'un concept de la taxonomie O_k est alors la suivante :

$$\lambda_k = \frac{PO_k - (\varepsilon_k * SO(O_k))}{|O_k|} \qquad [4]$$

Où :

- PO_k est le poids de la taxonomie O_k,
- $SO(O_k)$ est la somme des coefficients des concepts d'O_k,
- ε_k est la marge des poids d'un niveau $i+1$ par rapport au niveau i d'O_k,
- $|O_k|$ est le nombre de concepts dans la taxonomie O_k,

Exemple (suite). Le poids de base λ_1 et λ_2 d'un concept respectivement de la taxonomie O_1 et O_2 sont égaux à :

$$\lambda_1 = \frac{PO_1 - (\varepsilon * SO(O_1))}{|O1|} = \frac{0.571 - (0.01165 * 7)}{4} = 0.12236$$

$$\lambda_2 = \frac{PO_2 - (\varepsilon * SO(O_2))}{|O_2|} = \frac{1.429 - (0.0008 * 41)}{10} = 0.1396$$

Après avoir calculé le poids de base d'un concept, il s'agit à ce stade de calculer le poids effectif de chaque concept en tenant compte de la marge et de son coefficient, comme l'indique la Formule 5.

Le poids effectif d'un concept C_i de la taxonomie O_k, noté $PC(C_i,O_k)$, est calculé comme suit :

$$PC(C_i, O_k) = \lambda_k + (\varepsilon_k * Coeff(C_i, O_k)) \qquad [5]$$

Où :

- λ_k est le poids de base d'un concept de la taxonomie O_k,
- ε_k est la marge des poids d'un niveau $i+1$ par rapport au niveau i d'O_k.
- Coeff (C_i, O_k) est le coefficient du concept C_i d'O_k.

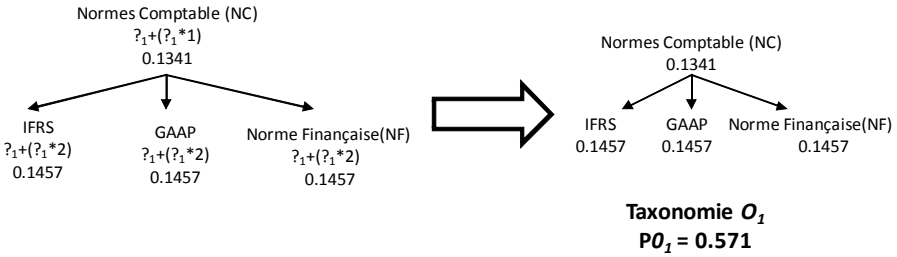
Notons que ces formules garantissent que la somme des poids de tous les concepts d'une taxonomie soit égale au poids de la taxonomie (soit à 1.429 pour O_2 de notre exemple). De plus, cette pondération diffère de celle utilisée dans la Figure 17 par le fait que les concepts feuilles auront plus d'importance que les concepts pères.

Exemple (suite). Le calcul des poids des concepts des taxonomies O_1 et O_2 de la Figure 18 donne :

$$PC(NC, O_1)=0,139+(0,0008*1)=0,1404$$
$$PC(IFRS, O_1)=0,139+(0,0008*4)=0,1428$$
$$PC(GAAP, O_1)=0,139+(0,0008*4)=0,1428$$
$$PC(NF, O_1)=0,139+(0,0008*2)=0,1412$$

$$PC(IS, O_2)=0,139+(0,0008*1)=0,1404$$
$$Coeff(FL, O_2)=0,139+(0,0008*4)=0,1428$$
$$Coeff(DB, O_2)=0,139+(0,0008*4)=0,1428$$
$$Coeff(DW, O_2)=0,139+(0,0008*2)=0,1412$$
$$Coeff(OLAP, O_2)=0,139+(0,0008*3)=0,142$$
$$Coeff(Storage, O_2)=0,139+(0,0008*5)=0,1436$$
$$Coeff(Design, O_2)=0,139+(0,0008*5)=0,1436$$
$$Coeff(Cube, O_2)=0,139+(0,0008*5)=0,1436$$
$$Coeff(,!O_2)=0,139+(0,0008*6)=0,1444$$
$$Coeff(Dimension, O_2)=0,139+(0,0008*6)=0,144$$

L'application de la Formule 5 sur les taxonomies O_1 *et* O_2 produit les poids présentés dans la Figure 19.



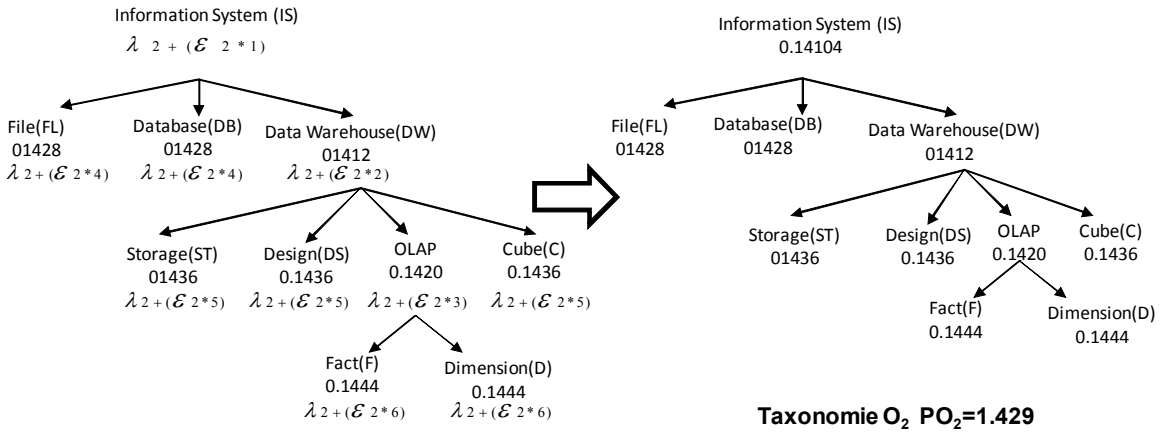


Figure 19. Poids des concepts des taxonomies O_1 et O_2 selon la Formule 5.

3.3.2 Choix d'une taxonomie pour un document

Selon notre approche, un entrepôt contient des documents appartenant à plusieurs domaines et dispose d'un ensemble de taxonomies de différents domaines. Dans cette phase, nous nous intéressons à déterminer une taxonomie pour chaque document. Il nous faut pour cela comparer la représentativité de chaque taxonomie par rapport au document, en pondérant les concepts des taxonomies en tenant compte du contenu du document. Nous pouvons ainsi déterminer le pouvoir représentatif de chaque taxonomie par rapport au document qu'elle va permettre d'indexer sémantiquement. Pratiquement, nous avons besoin de procéder aux calculs suivants :

- **Calcul du poids d'un concept pour un élément feuille du document**

Premièrement, nous calculons, pour chaque taxonomie de domaine candidate pour le document, le poids de chaque concept C_i par rapport à chaque élément feuille E_j du document *d*. Cette pondération est définie selon la Formule 6.

$$Poids\ (C_i, E_j) = \frac{\left| C_i^{E_j} \right|}{\left| C_i^{d} \right|} * PC\ (C_i, O_k) \qquad \forall\ j \qquad E_j \in d \qquad [6]$$

Où :

- $\left| C_i^{Ej} \right|$ est le nombre d'apparitions du concept C_i dans l'élément E_j,
- $\left| C_i^{d} \right|$ est le nombre d'apparitions de C_i dans le document *d*, et
- $PC(C_i,O_k)$ est le poids du concept C_i dans la taxonomie O_k.

Exemple (suite). Le résultat de l'application de la Formule 6 sur le document *XML-paper* avec les deux taxonomies O_1 et O_2 est schématisé dans la Figure 20. Les poids sont sur les lignes

en pointillés reliant les concepts de la taxonomie et les éléments du document. Par exemple, le poids du concept *OLAP* d'O_2 par rapport à l'élément *Name* de *Paper* est de 2,5. Alors que le poids de ce même concept par rapport à l'élément *Abstract/P1* est de 0,9.

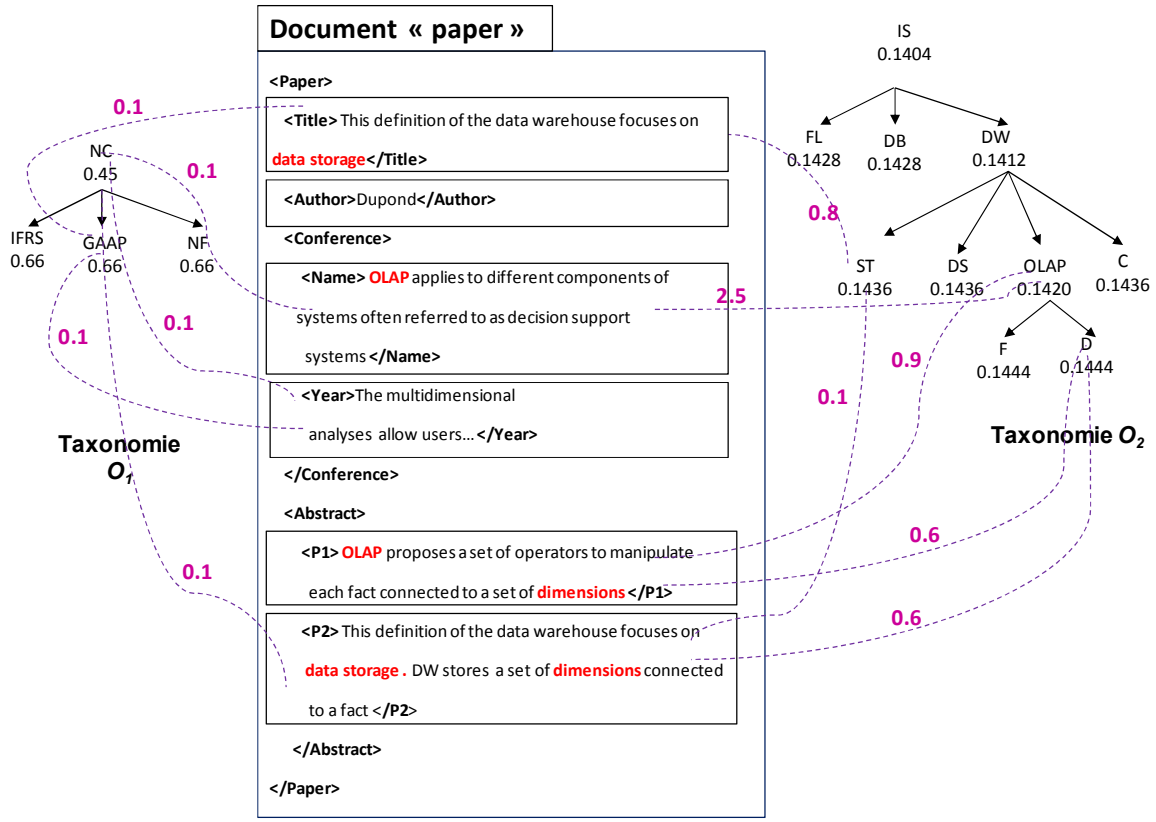


Figure 20. Poids des concepts taxonomiques par rapport aux éléments feuilles du document.

- **Calcul du poids d'un concept pour un document**

Ensuite, le poids de chaque concept C_i par rapport à tout le document d est calculé en fonction de la somme des poids de C_i dans les différents éléments de d, selon la Formule 7.

$$Poids(C_i, d) = \sum_{j=1}^{N} Poids(C_i, E_j) \qquad \forall\ j \qquad E_j \in d \qquad [7]$$

Où : N est le nombre d'éléments dans le document d.

Exemple (suite). Le résultat de l'application de la Formule 7 sur le document *XML-paper* avec les deux taxonomies O_1 et O_2 est montré dans la Figure 21. Les poids sont sur les lignes en pointillés reliant les concepts de la taxonomie et le document. Par exemple, le poids du concept *OLAP* de O_2 par rapport à *Paper* est de 3,4 (soit 0,9 + 2,5).

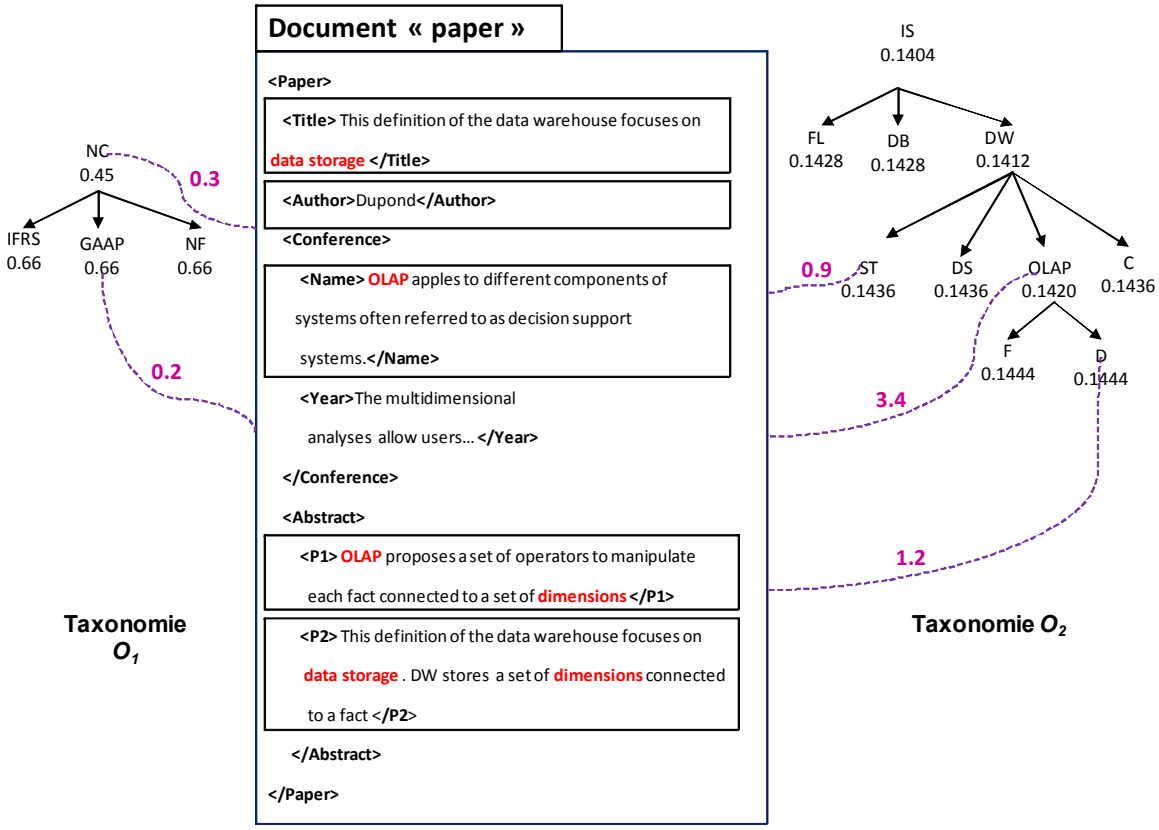


Figure 21. Poids des concepts taxonomiques par rapport au document.

- **Calcul du poids d'une taxonomie par rapport à un document**

Pour déterminer la taxonomie la plus représentative (i.e., appropriée) pour un document, il convient de calculer le poids de chacune des taxonomies candidates pour représenter la sémantique du document en question.

Le poids d'une taxonomie par rapport à un document correspond à la somme des poids des différents concepts appartenant à la taxonomie en question. Ce calcul est réalisé selon la Formule 8.

$$Poids\ (O_k\ ,\ d) = \sum_{i=1}^{|O_k|} Poids\ (C_i\ ,\ d) \quad [8]$$

Où :

- $|O_k|$ est le nombre de concepts de la taxonomie O_k,
- $Poids(C_i,d)$ est calculé selon la Formule 7.

Finalement, la taxonomie ayant le poids le plus élevé sera retenue pour déterminer la structure sémantique du document.

Exemple (suite). Le résultat de l'application de la Formule 8 sur le document *XML-paper* avec les deux taxonomies O_1 et O_2 est schématisé par la Figure 22.

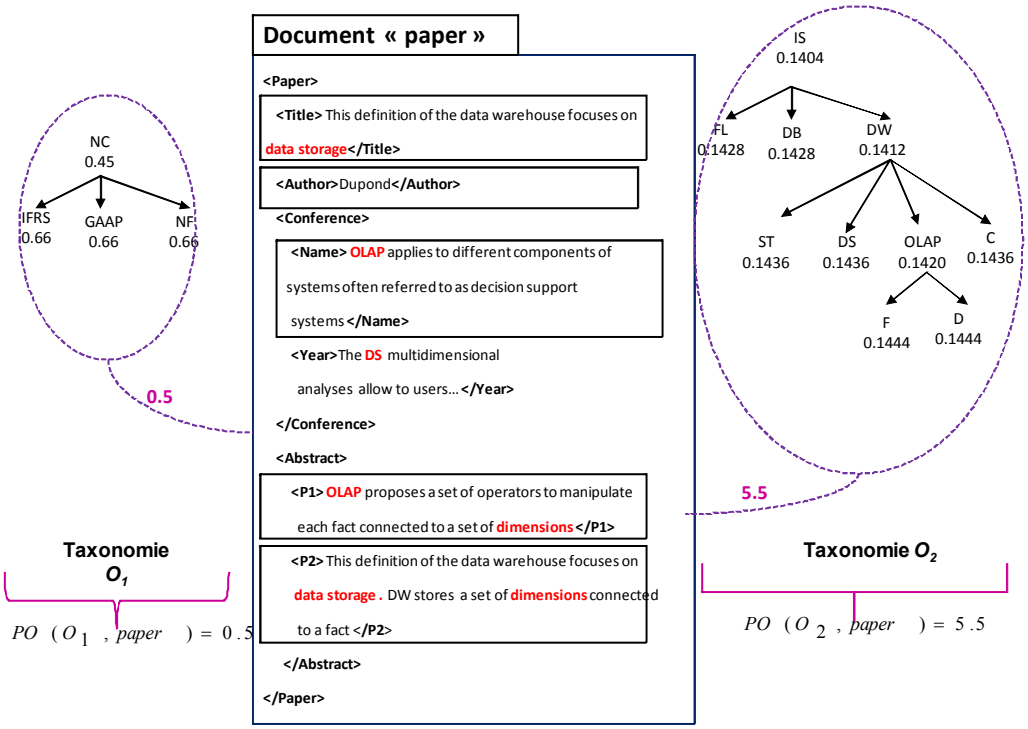


Figure 22. Poids de chaque taxonomie par rapport au document.

Exemple (suite). Ainsi, pour terminer, si nous considérons un document ayant la structure spécifique de la Figure 15, on aura le tableau suivant après application des Formules 6, 7 et 8.

Taxonomies		O_1				O_2										[F]
Eléments spécifique du document *paper*	Concept	NC	IFRS	GAAP	NF	IS	FI	DB	DW	ST	DS	OLAP	C	F	D	
	Title	0	0	0.1	0	0	0	0	0	0.8	0	0	0	0	0	[6]
	Author	0	0	0	0	0	0	0		0	0	0	0	0	0	
	Name	0.1	0	0	0	0	0	0	0	0	0	2.5	0	0	0	
	Year	0.1	0	0.1	0	0	0	0	0	0	0	0	0	0	0	
	P1	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0.6	
	P2	0.1	0	0	0	0	0	0	0	0.1	0	0	0	0	0.6	
$Poids(C_i, d)$		0.3	0	0.2	0	0	0	0	0	0.9	0	3.4	0	0	1.2	[7]
$Poids(O_k, d)$		0.5				5.5										[8]

Tableau 4. Poids des concepts pour le choix d'une taxonomie.

Finalement, pour le document *XML-paper*, le poids de la taxonomie O_2 (5.5) est supérieur au poids de la taxonomie O_1 (0.5). La taxonomie O_2 sera alors retenue pour construire la structure sémantique du document *XML-paper*.

3.4 Affectation des concepts aux éléments feuilles

L'objectif de cette phase est d'affecter un seul concept représentatif à chaque élément feuille du document, et ceci en se basant sur les poids des concepts calculés par la Formule 4.

Pour un élément feuille E_j, quatre cas se présentent :

- **Cas 1** : Aucun concept n'a pu être déterminé pour E_j. Nous introduisons dans ce cas le concept hypothétique vide appelé *Null* que nous affectons à l'élément E_j (cf. Figure 23).

– **Cas 2** : Un seul concept est déterminé pour E_j ; il sera retenu et *affecté à* E_j comme concept représentatif (cf. Figure 24).

– **Cas 3** : Plusieurs concepts déterminés à partir de l'élément feuille E_j appartiennent à une même hiérarchie dans la taxonomie du document (cf., Figure 25). Dans ce cas, nous prévoyons deux situations :

  – Si les poids calculés pour ces concepts sont très proches (moyennant un seuil arbitraire fixé par l'utilisateur), nous affectons à E_j le concept le plus spécifique dans la hiérarchie.

  – Autrement, si les poids de ces concepts sont divergents, nous affectons à E_j le concept ayant le poids le plus élevé, indépendamment de sa position dans la hiérarchie.

– **Cas 4** : Plusieurs concepts sont déterminés pour E_j et appartiennent à plusieurs hiérarchies dans la taxonomie (cf. Figure 26). Dans ce cas :

  – Si les poids des concepts sont très proches (au sens précédent), nous affectons à E_j le concept le plus spécifique des deux concepts dans la taxonomie.

  – Si les poids des concepts sont divergents, nous affectons à E_j le concept ayant le poids le plus élevé.

Exemple (suite). Les Figures 23 à 26 illustrent les différents cas d'affection de concepts aux éléments feuilles de la structure logique du document *XML-paper*.

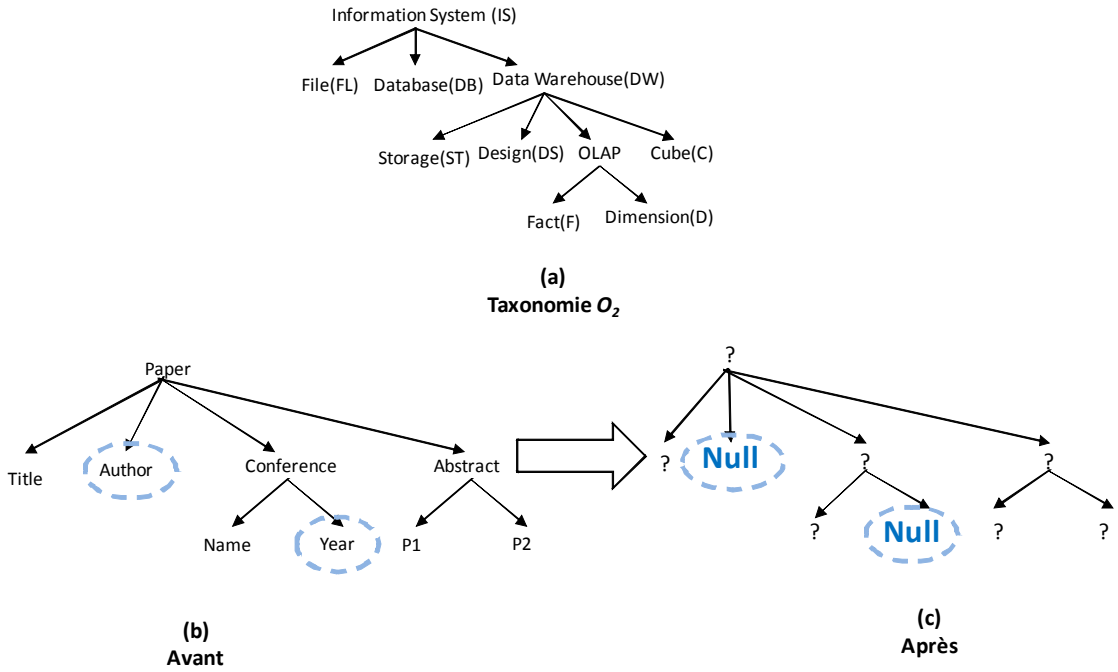


Figure 23. Cas 1 : Affectation du *Null* aux éléments feuilles pour lesquels aucun concept n'a été déterminé.

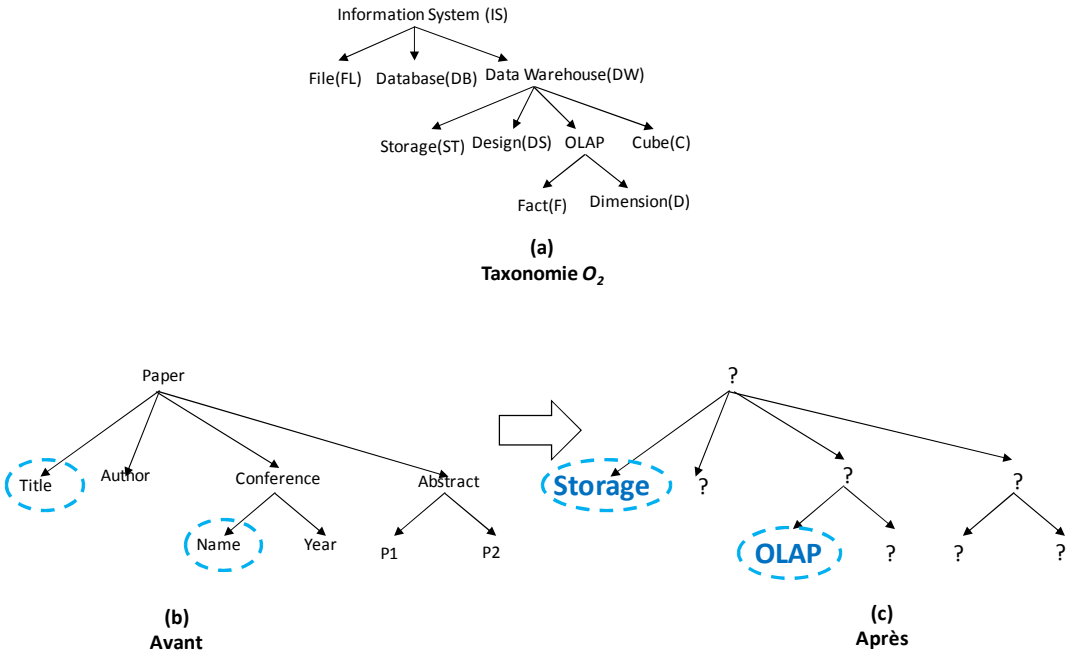


Figure 24. Cas 2 (concept unique). Affectation d'un concept à un élément feuille.

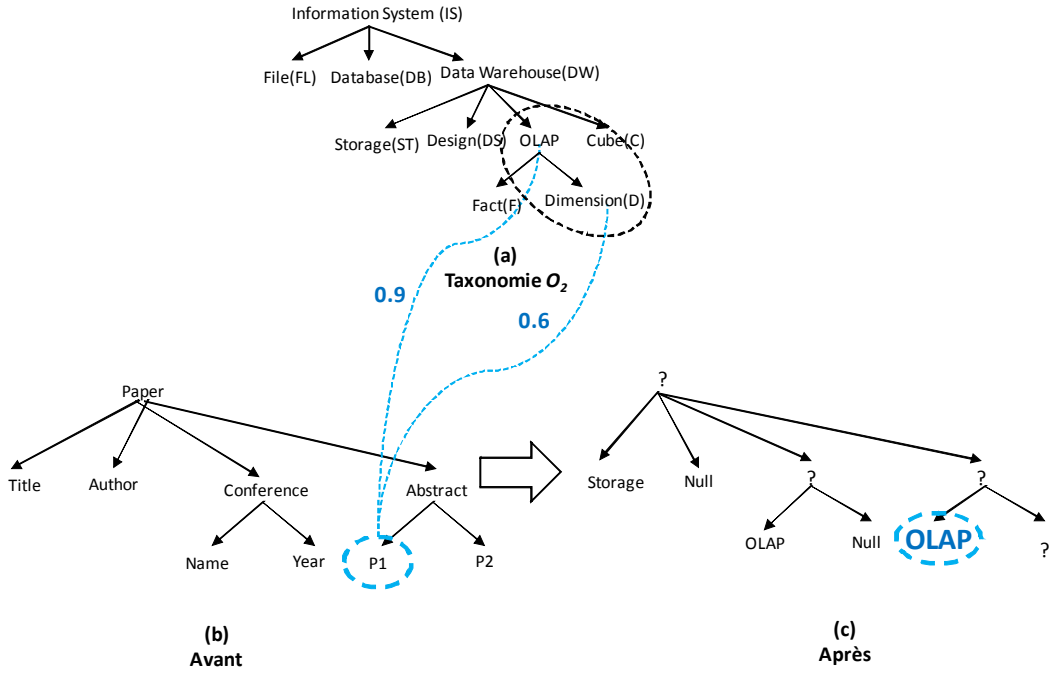


Figure 25. Cas 3 : Affectation d'un concept parmi plusieurs d'une même hiérarchie à un élément feuille.

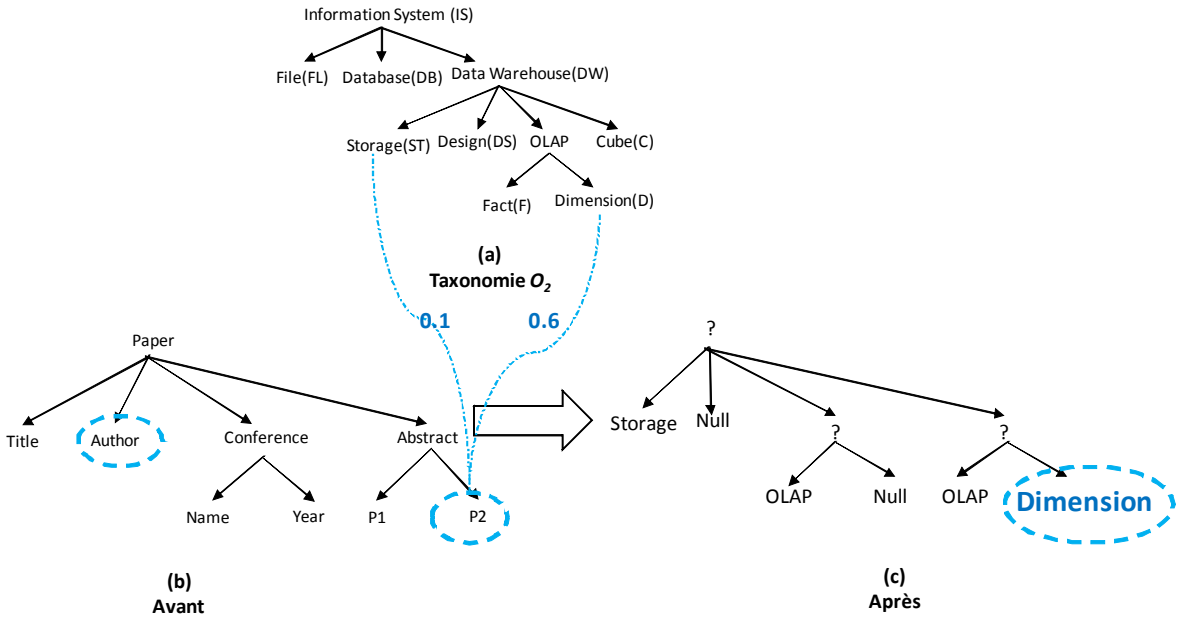


Figure 26. Cas 4 : Affectation d'un concept parmi plusieurs appartenant à des hiérarchies différentes à un élément feuille.

3.5 Méthode d'inférence des concepts

Jusqu'à présent, nous avons déterminé, pour chaque élément feuille de la structure sémantique d'un document, un concept choisi à partir de la taxonomie retenue pour le

document, ou bien le concept *Null* à défaut. Nous poursuivons notre objectif afin de finaliser la construction de la structure sémantique.

Il s'agit maintenant d'assigner des concepts aux nœuds non feuilles de la structure. Du fait que ces nœuds n'ont pas d'information textuelle associée, nous procédons alors par inférence des concepts des feuilles vers leurs ascendants en appliquant les règles suivantes :

- **Règle 1** : Si un nœud père possède un seul fils alors il aura le même concept que son fils (cf. Figure 27).
- **Règle 2** : Si un nœud père possède plusieurs fils dont les concepts appartiennent à une même hiérarchie de la taxonomie, alors ce père aura le concept le plus générique des concepts associés à ses fils (cf. Figure 28).
- **Règle 3** : Si un nœud père possède plusieurs fils dont les concepts appartiennent à plusieurs hiérarchies de la taxonomie, alors le concept attribué à ce père est l'ancêtre commun des concepts associés à ses fils (cf. Figure 29).

Exemple (suite). Les Figures 27 à 29 illustrent les différents cas d'inférence de concepts des nœuds feuilles vers leur ascendant.

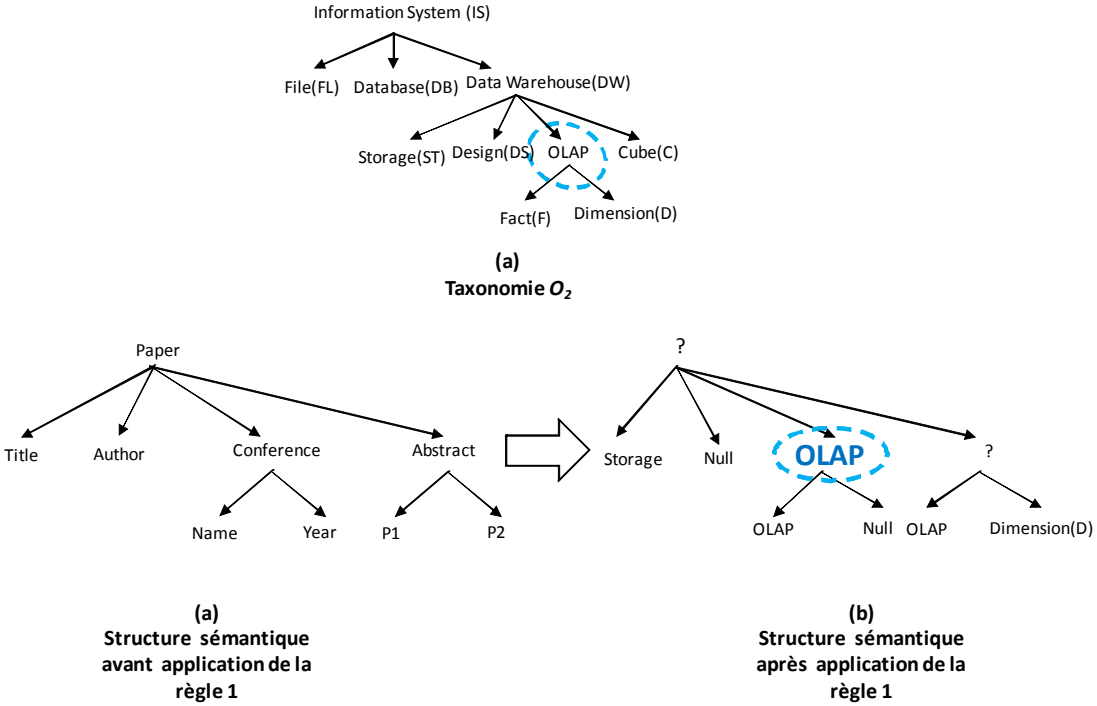


Figure 27. Illustration de l'inférence de concepts : Règle 1.

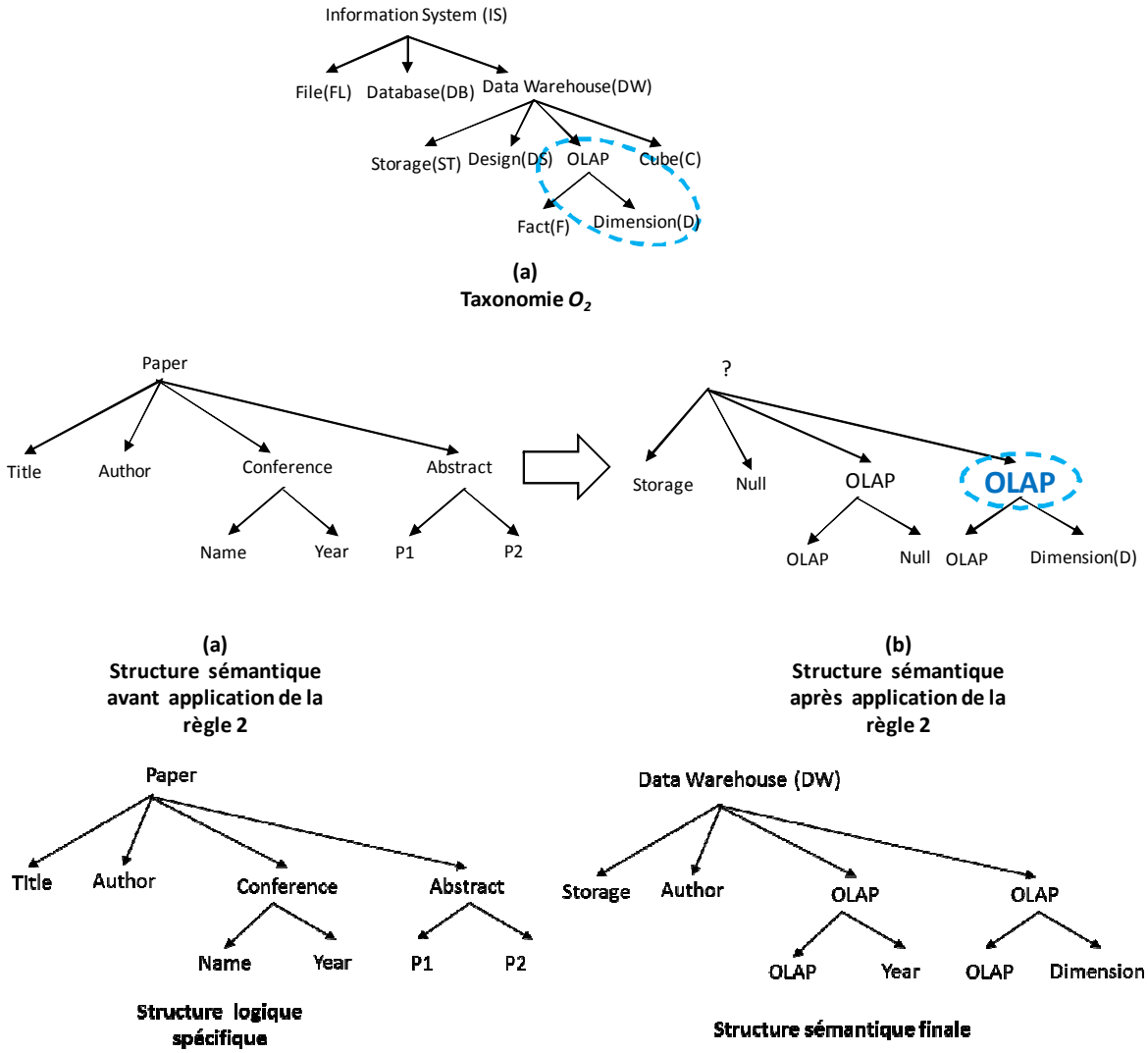


Figure 28. Illustration de l'inférence de concepts : Règle 2.

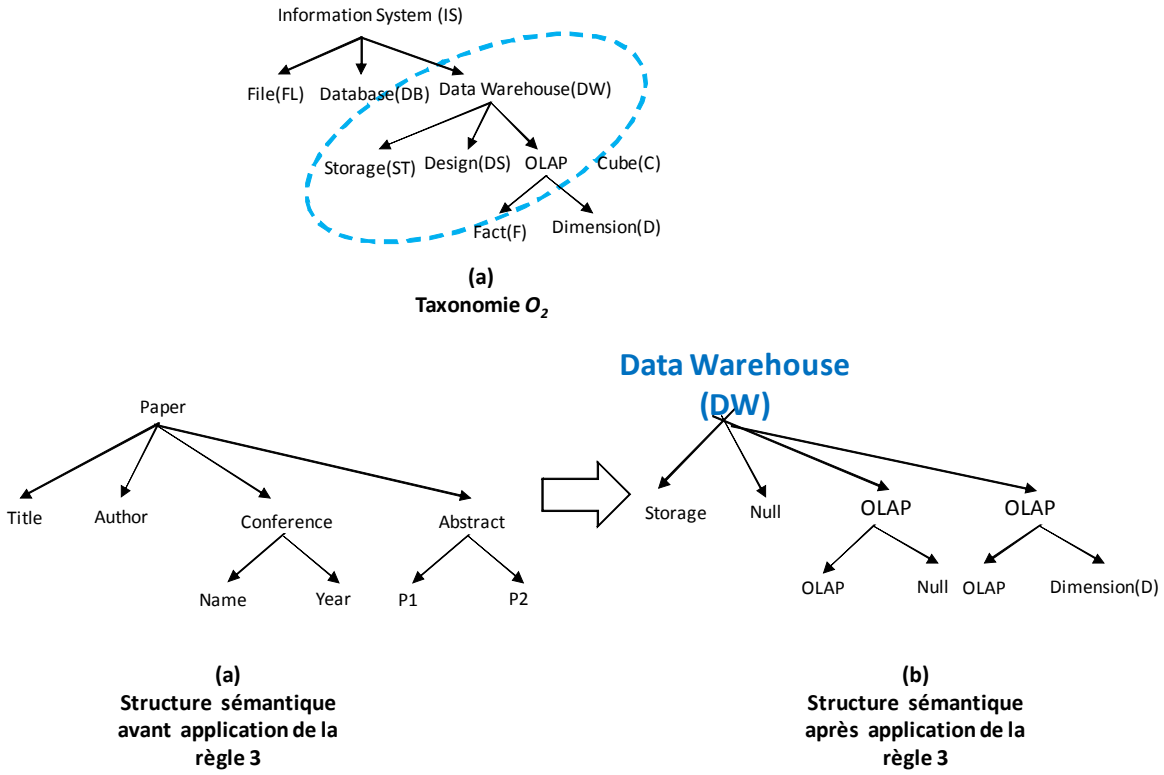


Figure 29. Illustration de l'inférence de concepts : Règle 3.

Par exemple, dans la Figure 29, le concept associé au nœud racine est l'ancêtre commun de *Storage* et *OLAP* de la taxonomie, qui est *Data Warehouse.*

3.6 Affectation des métadonnées aux éléments sans concepts

A l'issue de l'étape d'inférence, tous les éléments de la structure spécifique du document sont associés soit à des concepts de la taxonomie retenue, soit au concept *Null*.

La dernière étape de la démarche de détermination des concepts à associer aux éléments de la structure logique consiste à identifier, pour les éléments de la structure logique spécifique non remplacés par des concepts (i.e. ayant reçu le concept *Null*), les labels des métadonnées correspondantes. Ces labels de métadonnées représenteront alors l'élément dans la structure sémantique.

Exemple (suite) : La structure sémantique finale du document "XML-paper" obtenue après déroulement des 5 phases est celle présentée dans la Figure 30. Les deux éléments *Author* et *Year* de la structure logique qui s'étaient vus associer le concept *Null* (cf. Figure 23, Cas 1 de la démarche), auront pour concept de label de la métadonnée associé, i.e. *Author* et *Year*.

Figure 30. Structures, logique spécifique et sémantique, du document "XML-paper ".

3.6 Conclusion

Dans ce chapitre, nous avons présenté notre approche qui permet de définir une structure sémantique pour un document XML orienté-texte à partir de sa structure logique spécifique, de son contenu et en recourant à une ressource sémantique de domaine (une taxonomie dans notre cas). Dans cette approche, nous avons commencé par l'extraction des termes significatifs des éléments feuilles (fragments textuels) d'un document XML. Ensuite, nous avons sélectionné une taxonomie (parmi un ensemble de taxonomies candidates) que nous avons associée au document ; il s'agit de la taxonomie qui correspond le mieux à la sémantique du document. Finalement, nous avons associé à chaque élément de la structure spécifique du document, un concept significatif issu de la taxonomie retenue. Pour ce faire, nous avons défini une démarche d'affectation des concepts basée sur 7 formules, 4 cas d'affectation des concepts aux éléments feuilles, et 3 règles d'inférence des concepts.

Il est important de noter que ces structures sémantiques pourront être utilisées dans divers domaines et à plusieurs titres. Elles permettront en particulier d'enrichir les analyses multidimensionnelles de documents dans l'entrepôt. Une description d'analyse multidimensionnelle basée sur cette structure sémantique est présentée dans (Khrouf et al., 2011).Elles pourront également intervenir dans des processus de Recherche d'Information ou de Gestion Électronique de Documents pour lesquels les aspects sémantiques apportent énormément en matière de désambiguïsation et d'augmentation de la pertinence lors des recherches. En effet, la combinaison en particulier d'une structuration logique et d'une structuration sémantique est un facteur important d'amélioration de la pertinence et de la précision des traitements en permettant de se focaliser sur des parties spécifiques de documents (et non plus uniquement sur les documents dans leur totalité).

Dans le chapitre suivant, nous nous intéressons à l'exploitation de la structure sémantique des documents XML, d'abord en intégrant ces structures dans l'entrepôt de documents et, ensuite en les exploitant dans les interrogations à travers le langage de requêtes XQuery.

Chapitre 4 : Exploitation de la structure sémantique

Sommaire

4.1 Introduction

La structure sémantique d'un document reflète l'organisation de la sémantique contenue dans les éléments textuels de ce document. Elle est définie au travers de la composition sémantique, représentant le sens d'un ou de plusieurs éléments de la structure logique (Poullet L., 1997).

L'utilité de la structure sémantique réside dans la manipulation du document selon des propriétés sémantiques particulières à l'information contenue dans le document ; « la sémantique est définie par la composition d'objets sémantiques décrivant le sens des nœuds feuilles ou non feuilles de la structure logique. Les objets sémantiques sont en outre typés en fonction du but rhétorique des objets logiques qui leurs sont associés » (Poullet L., 1997). Dans ce contexte, nous proposons deux contributions, premièrement l'intégration de la structure sémantique à l'entrepôt de documents, et deuxièmement l'interrogation de l'entrepôt basée sur cette structure. Pour ce faire, nous étendons le méta-modèle de l'entrepôt de documents de (Khrouf K., 2004) pour supporter les éléments de la structure sémantique. Ensuite, nous proposons un processus pour une interrogation qui tient compte des éléments sémantiques décrivant les documents XML et ceci par le biais du langage de requêtes XQuery. Pour cette interrogation, nous définissons un *méta-document* qui a pour but de décrire l'organisation logique et sémantique d'un document.

Dans ce qui suit, nous commençons par l'intégration de la structure sémantique dans l'entrepôt de documents. Ensuite, nous introduisons les langages de requêtes spécifiques aux documents XML, et en particulier le langage XQuery. Enfin, nous proposons un ensemble de requêtes d'interrogation des documents XML dans un contexte de Recherche d'Information (RI) et dans un contexte d'analyse multidimensionnelle OLAP, requêtes qui exploitent la structure sémantique de ces documents.

4.2 Intégration de la structure sémantique à l'entrepôt de documents

Dans les travaux visant à la modélisation et l'entreposage de documents, (Khrouf K., 2004) a proposé un méta-modèle de documents (cf. Figure 31) permettant le regroupement des documents selon des structures identiques ou approchantes afin d'appliquer les techniques d'analyses multidimensionnelles aux informations documentaires.

4.2.1 Méta-modèle d'entrepôts de documents

Ce méta-modèle décrit et intègre les composants suivants :

1. Les documents intégrés dans l'entrepôt.
2. Une partie structurelle qui décrit la structure logique hiérarchique d'un document. Nous distinguons deux structures :

a) Un modèle générique (cf. Figure 31-b) qui contient une classe *structure générique* nommée *Str_Gen* commune à un ensemble de documents. Cette structure est définie par :

- Un ensemble d'*éléments génériques* (classe *Elts_Gen*) pouvant être composés d'autres éléments génériques (*Elts_Gen*) dont chacun est caractérisé par un nom unique, une cardinalité et éventuellement des éléments ou des attributs génériques.
- Des *attributs génériques* (classe *Atts_Gen*). Un attribut décrit un élément générique, c'est-à-dire que cet attribut ajoute des informations concernant l'élément en question.

b) Un modèle spécifique (cf. Figure 31-c) conforme au modèle générique de la Figure 31-b ; il décrit la structure spécifique propre à un document. Ce modèle est défini par :

- Un ensemble d'*éléments spécifiques* (classe *Elts_Spec*), chaque élément est caractérisé par un numéro de séquence (ordre d'apparition de l'élément spécifique pour l'élément générique considéré).
- Des *attributs spécifiques* (classe *Atts_Spec*), chacun décrit une information concernant un élément spécifique. Cet attribut doit correspondre obligatoirement à un attribut générique.

3. La partie contenu (instances) : il s'agit de la description du contenu textuel des éléments de la structure spécifique représentée dans le modèle (cf. Figure 31-d).

Le diagramme de classes UML de la Figure 31 modélise ces différents composants :

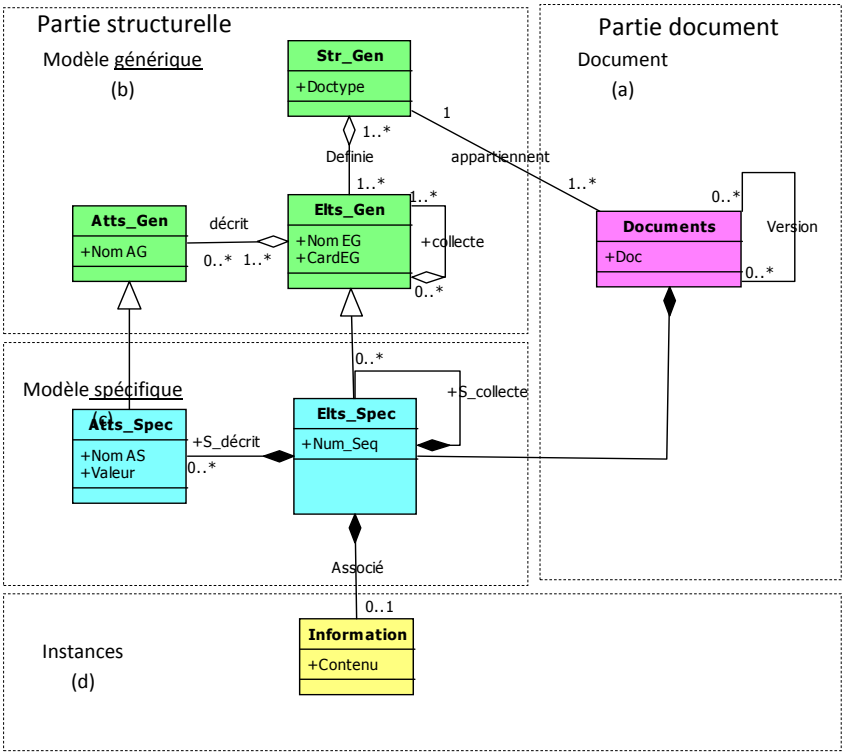


Figure 31. Méta-modèle d'un entrepôt de documents (diagramme de classess UML) (Khrouf K. et al., 2012) .

Ce méta-modèle permet :

- la gestion de l'hétérogénéité de structures,
- la mise en évidence des différents niveaux de granularité des documents,
- le regroupement des documents selon des structures communes identiques ou approchantes.

Dans ce qui suit, nous présentons un exemple d'instanciation de ce méta-modèle. Soit la structure générique intitulée *paper* définie par les éléments génériques suivants : *Title*, *Author*, *Conference et Abstract.* L'élément *Conference* est composé de deux éléments génériques *Name* et *Year*.

La Figure 32 décrit la structure générique *Paper* selon le formalisme *XML Schéma*.

```
<xs :schema xmlns :xs="http ://www.w3.org/2001/XMLSchema" elementFormDefault="qualified"
attributeFormDefault="unqualified">
        <xs:element name="paper" >
        <xs:complexType >
        <xs:sequence>
                <xs:element name="title"    type= "xs:string"/>
                <xs:element name="Author"    type= "xs:string"/>
                <xs:element name="Conference"   >
<xs:complexType>
<xs:sequence>
                <xs:element name="Name"    type= "xs:string"/>
                <xs:element name="Year"    type= "xs:string"/>
                </xs:sequence>
                </xs:complexType>
                </xs:element>
        <xs:element name="Abstract"   type= "xs:string"/>
        </xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>
```

Figure 32. Schéma XML de la structure générique *Paper*.

Soient les deux documents *Doc1.XML* et *Doc2.XML* de la Figure 33, à intégrer dans l'entrepôt, et conformes à la structure générique *Paper*. L'instanciation de ces deux documents se fait comme indiqué dans la Figure 34.

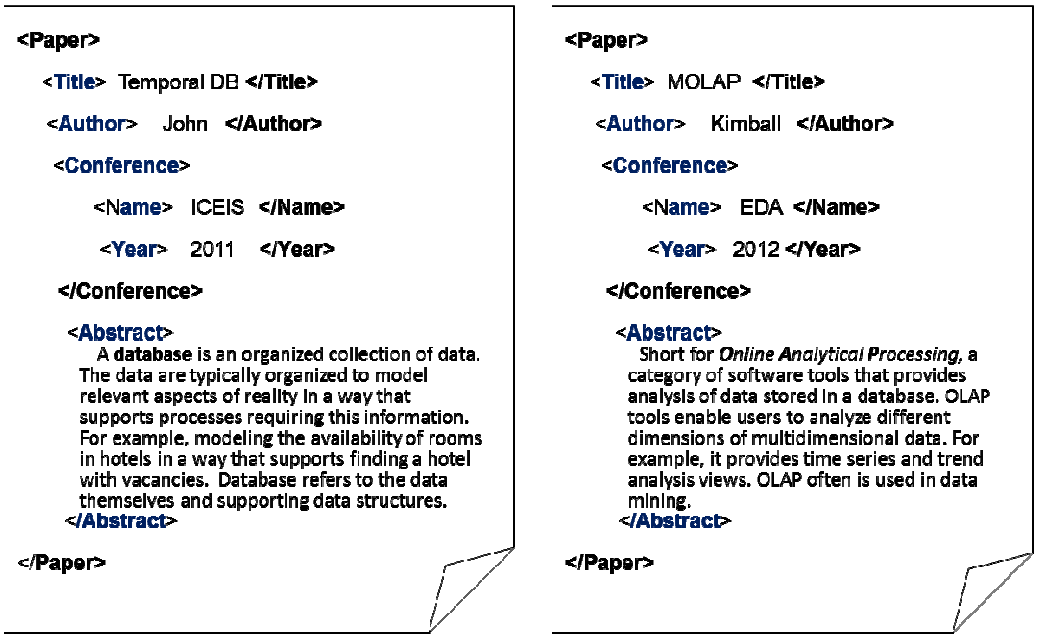


Figure 33. Deux exemples de documents XML.

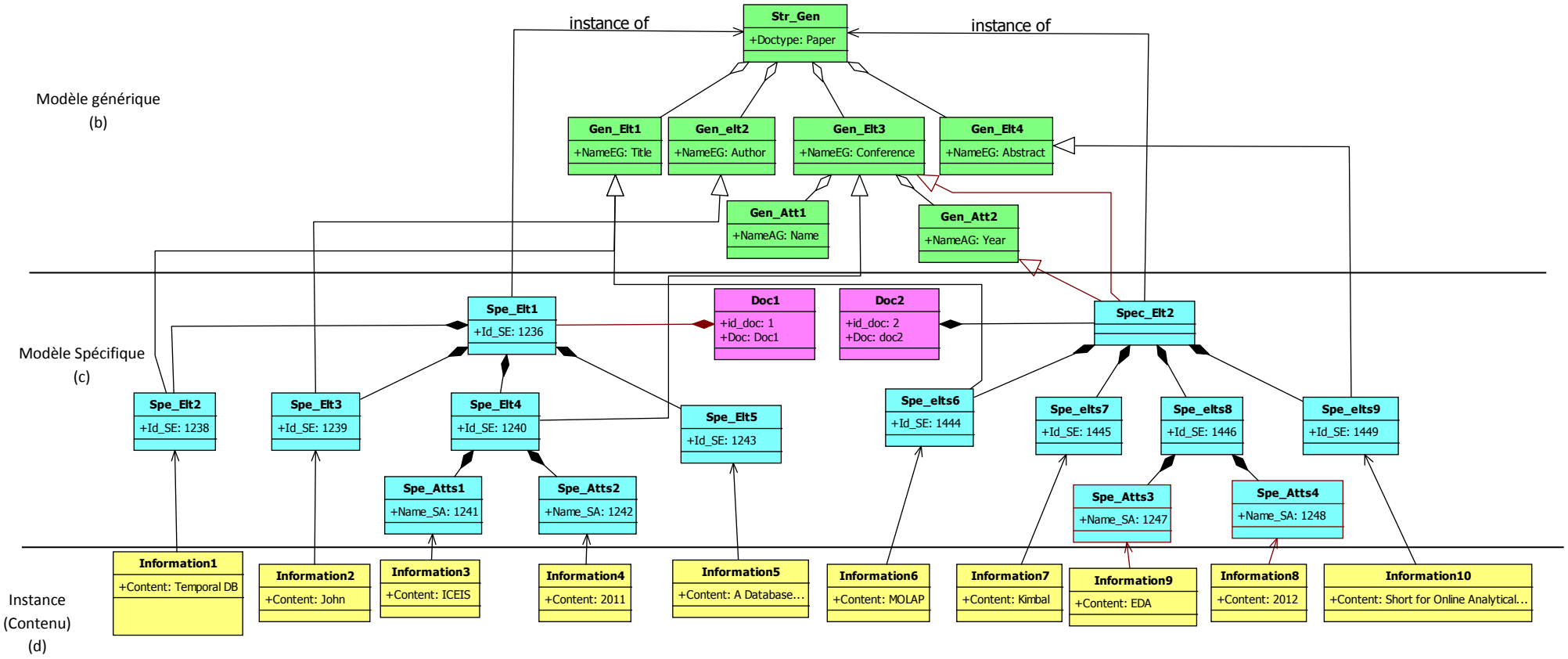


Figure 34. Instanciation du méta-modèle par les deux documents *Doc1.XML* et *Doc2.XML*.

- *Doc1.XML* a pour titre (1238, **Temporal *DB***), pour auteur (1239, **John**), est associé à la conférence (1240) possédant comme nom (1241, ***ICEIS***) qui s'est déroulée les années (1242, **2011**), et a pour résumé (1243, "**A database is…**").
- *Doc2.XML* a pour titre (1444, ***MOLAP***), comme auteur (1445, **Kimball**), est associé à la conférence (1246) possédant le nom (1247, ***EDA***) qui s'est déroulée l'année (1248, **2012**), et a pour résumé (1449, "**Short for…**").

En se basant sur ce méta-modèle, (Khrouf K. et al., 2005) a proposé une démarche pour analyser d'une manière multidimensionnelle les informations factuelles des documents XML intégrés dans l'entrepôt. Cependant, les techniques d'analyse proposées s'appliquent essentiellement à la structure générique des documents constitués, généralement, par des éléments factuels courts et précis (*titre, auteur, année*, etc.), et par conséquent, elles ne peuvent pas être appliquées sur des éléments textuels de grande taille (*section*, *paragraphe*, etc.). Ceci représente un handicap de taille pour les analyses décisionnelles qui, désormais, auront besoin d'explorer le contenu informationnel ; c'est-à-dire en exploitant la sémantique du document.

Afin d'améliorer ces travaux, nous procédons à une extension de la démarche d'analyse multidimensionnelle pour tenir compte de la sémantique des documents. Pour ce faire, cette extension suppose que chaque élément de la structure spécifique d'un document puisse être projeté sur un concept d'une ressource sémantique telle qu'une taxonomie de domaine. Ainsi, l'objectif de ce travail est : (a) d'une part d'étendre le méta-modèle de la Figure 31 de manière à affecter à chaque document de l'entrepôt une taxonomie ; (b) et, d'autre part, d'identifier le concept taxonomique à associer à chaque élément de la structure logique d'un document (section, paragraphe, etc.).

4.2.2 Extension du méta-modèle d'entrepôts de documents

Afin de pouvoir réaliser des analyses multidimensionnelles en se basant sur le contenu textuel des documents en plus de leur structure, nous proposons une extension du méta-modèle de la Figure 31.

4.2.2.1 Description du méta-modèle étendu

Ce méta-modèle sera enrichi par une partie dédiée à la sémantique. Cette partie est définie par un ensemble de taxonomies. En se basant sur les résultats du chapitre 3, chaque document de l'entrepôt sera associé à une taxonomie (lien entre classes *Documents* et *Taxonomy* (cf. Figure 35). Les éléments de la structure logique seront décrits par les concepts correspondants (lien entre la classe *Information* et la classe *Concepts*). Il s'agit de chercher, dans la taxonomie de domaine associée au document en question, le concept le plus approprié, c'est-à-dire le concept qui décrit la sémantique de l'élément feuille et ceci en tenant compte des mots-clés contenus dans le texte de cet élément. Le concept trouvé est alors assigné à cet élément feuille.

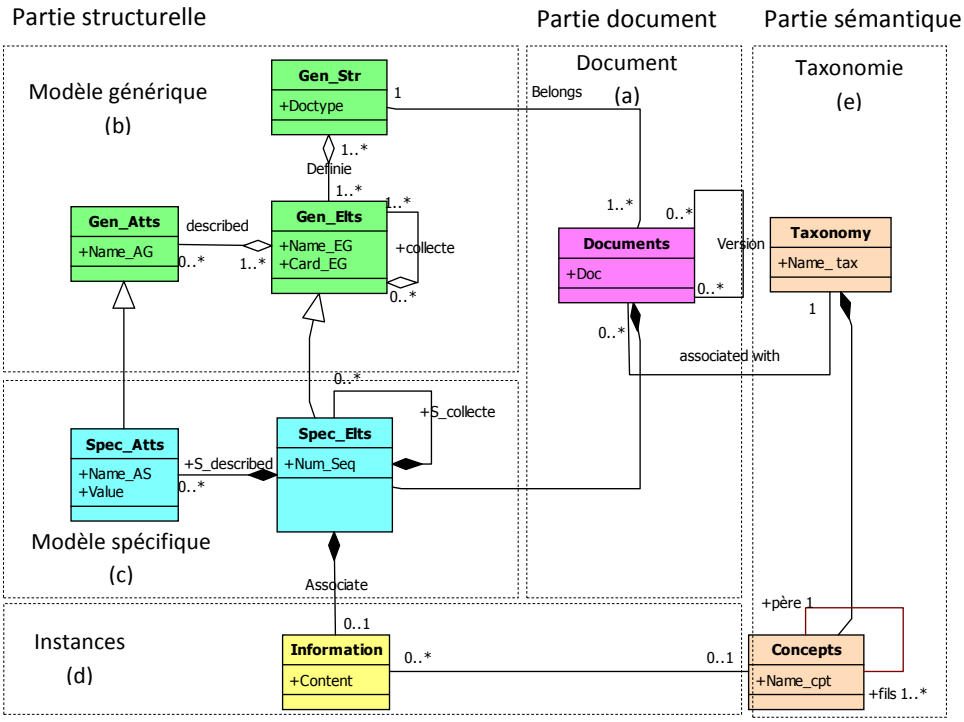


Figure 35. Méta-modèle étendu d'un entrepôt de documents intégrant la structure sémantique (formalisme diagramme de classess UML) (Ben Meftah S. et al., 2013).

4.2.2.2 Exemple d'instanciation

Dans l'exemple d'instanciation suivant (cf. Figure 36), nous reprenons l'exemple de la Figure 34 et nous ajoutons la partie sémantique. Ainsi, les deux documents *Doc1.XML* et *Doc2.XML* sont associés à la taxonomie *Information System* "*IS*" (cf. Chapitre 3, section 3.3.1) et leurs éléments textuels aux trois concepts *IS (Information System), DB (DataBase), DW (Data Warehouse)* et *OLAP* de cette taxonomie.

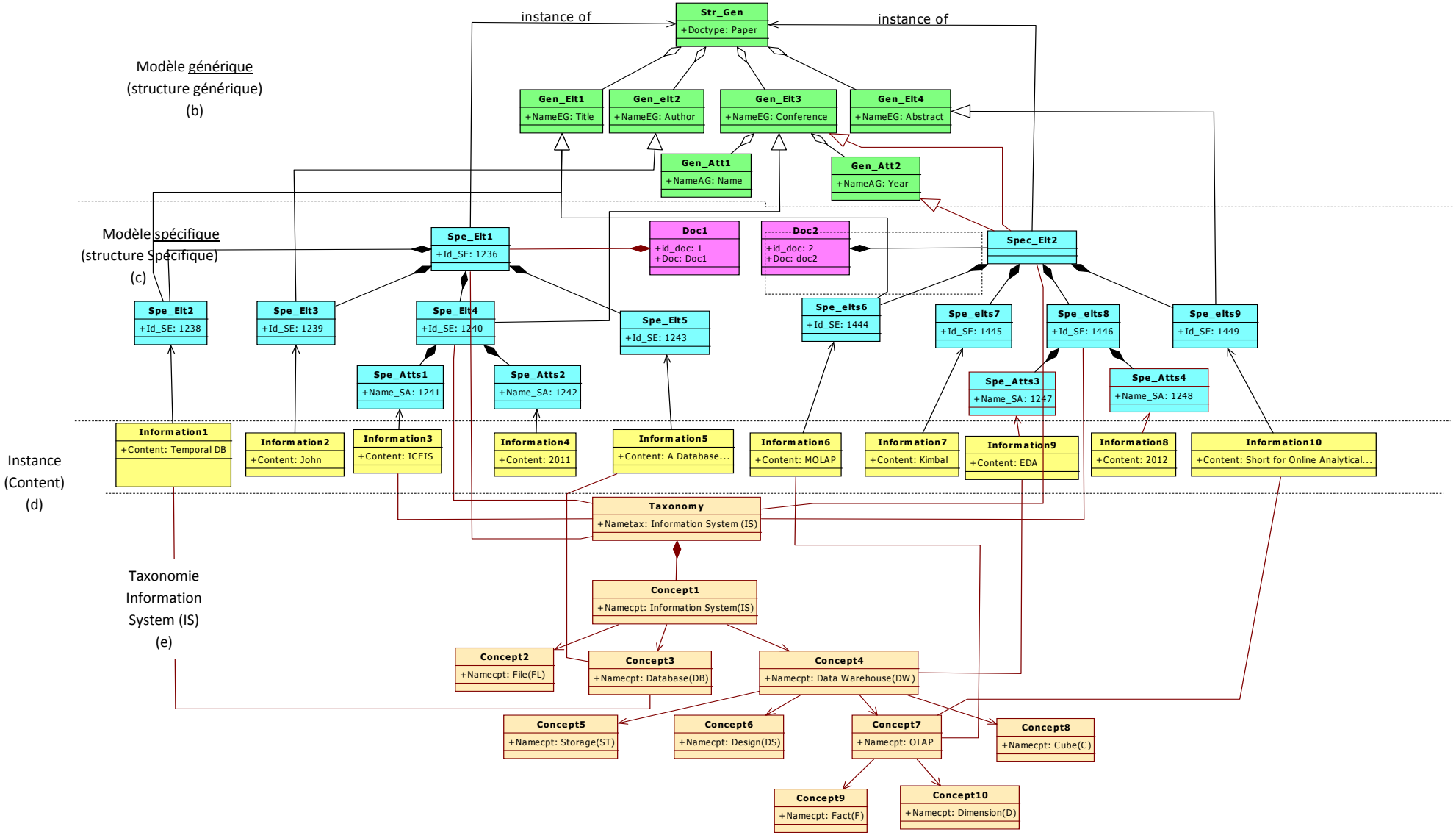


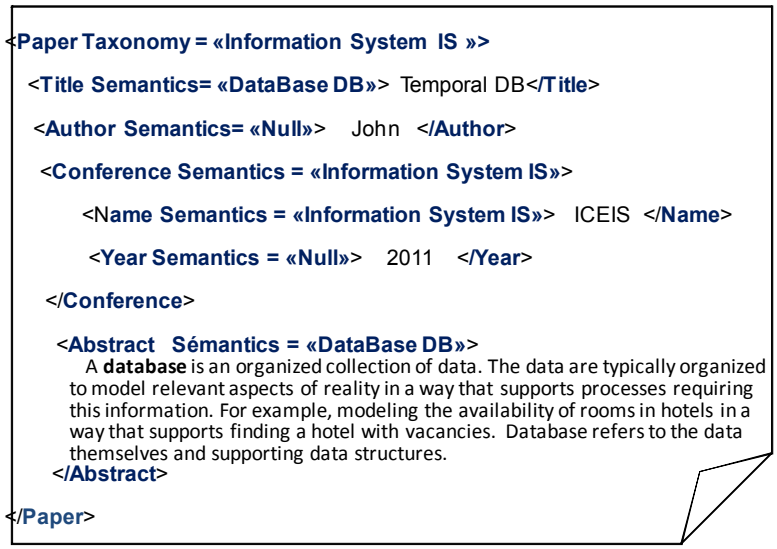
Figure 36. Exemple d'instanciation du méta-modèle étendu pour les documents *Doc1.XML et Doc2.XML.*

4.3. Concept de Méta-document

Dans ce travail, nous souhaitons aussi interroger les structures sémantiques, notamment sans passer par le contenu des documents de l'entrepôt et ceci pour des raisons de performanes. A cette fin, nous devons stocker les structures sémantiques construites. Deux solutions se présentent :

- La première solution consiste à intégrer cette structure sémantique dans le document XML. Pour ce faire, nous pouvons ajouter un attribut *Taxonomy* dans la balise racine du document et une balise nommée *Semantics* dans les autres balises.

Exemple : la Figure 37 montre le document Doc1.xml de la Figure 33 après modification.



Doc2 modifié.XML

Figure 37. Document *Doc1.XML* enrichi par la sémantique des balises selon la première solution

- La deuxième solution consiste à construire pour chaque document un autre document qui décrit la structure sémantique du document XML en question ; nous l'appelons *méta-document*.

Nous avons opté pour cette deuxième solution pour les trois raisons suivantes :

(1) Éviter de modifier les documents XML source (propriété intellectuelle),

(2) Assurer la transparence des méta-documents aux utilisateurs afin d'éviter de compliquer l'interrogation.

(3) Assurer l'indépendance des documents initiaux par rapport à leur structure sémantique en cas de modification de cette dernière.

La Figure 38 présente le processus que nous avons proposé pour l'interrogation sémantique de l'entrepôt de documents.

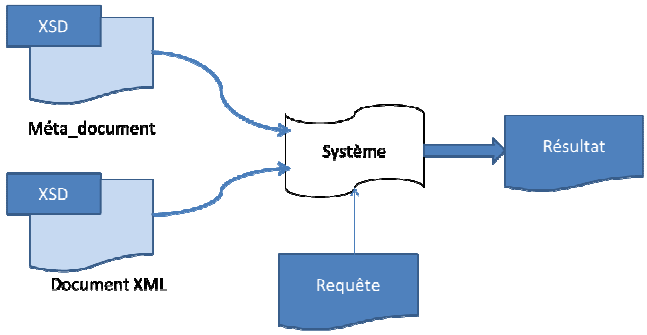


Figure 38. Interrogation sémantique des documents.

Nous décrivons le méta-document par un ensemble de balises dont chacune correspond à un nœud de la structure logique du document et possède les attributs suivants :

- **Nom** : nom de la balise dans la structure logique ;
- **Contenu** : référence (lien Xlink) vers le contenu textuel du nœud que décrit la balise en question ;
- **Sémantique** : contient le concept associé à la balise.

Exemple : La Figure 39 montre le méta-document associé à la structure logique du document *Doc2.XML* de la Figure 33.

```
<?xml version="1.0" encoding="UTF-8"?>
<Balises>
<Balise
      Nom ="Paper"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper'">
      semantique ="Information System IS/">
<Balise
       Nom="Title"
       Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Title'"
       semantique ="DataBase DB"/>
<Balise
      Nom="Author"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Author'"
      semantique="NULL"/>
<Balise
      Nom="Conference"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Conference'"
      semantique="Information System IS" />
<Balise
      Nom="Name"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Conference\Name'"
      semantique="Information System IS"/>
<Balise
      Nom="Year"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Conference\Year'"
      semantique="NULL"/>
<Balise
      Nom="Abstract"
      Contenu ="xlink:href= 'C:\Doc2.xml\Paper\Abstract"
      semantique="DataBase DB"/>
</Balise>
</Balises>
```

Figure 39. Méta-document associé au document *Doc2.XML*

La Figure 40 correspond à la structure du méta-document exprimée selon le formalisme standard XML Schéma qui permet de valider les documents XML.

```
<?<!--W3C Schema generated by XMLSpy v2009 sp1 (http://www.altova.com)-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
	<xs:element name="Paper">
	<xs:complexType>
		<xs:attribute name="semantics" use="required">
			<xs:simpleType>
			<xs:restriction base="xs:string">
				<xs:enumeration value="Information System IS"/>
			</xs:restriction>
		</xs:simpleType>
		</xs:attribute>
		<xs:attribute name="contenue" use="required">
			<xs:simpleType>
			<xs:restriction base="xs:string">
				<xs:enumeration value="xlink:href= &apos;C:\Doc2\Paper.xml&apos;"/>
			</xs:restriction>
			</xs:simpleType>
		</xs:attribute>
		<xs:attribute name="Name" use="required">
			<xs:simpleType>
			<xs:restriction base="xs:string">
				<xs:enumeration value="titre"/>
			</xs:restriction>
			</xs:simpleType>
		</xs:attribute>
	</xs:complexType>
	</xs:element>
	<xs:element name="record">
	<xs:complexType>
	<xs:sequence>
	<xs:element ref="title"/>
	</xs:sequence>
	<xs:attribute name="semantics" use="required">
		<xs:simpleType>
		<xs:restriction base="xs:string">
			<xs:enumeration value="DataBase DB"/>
		</xs:restriction>
		</xs:simpleType>
	</xs:attribute>
	<xs:attribute name="contenue" use="required">
		<xs:simpleType>
		<xs:restriction base="xs:string">
			<xs:enumeration value="xlink:href= &apos;C:\ Doc2\Paper.xml&apos;"/>
		</xs:restriction>
		</xs:simpleType>
	</xs:attribute>
	<xs:attribute name="Name" use="required">
	<xs:simpleType>
		<xs:restriction base="xs:string">
		<xs:enumeration value="Auteur"/>
		</xs:restriction>
	</xs:simpleType>
	</xs:attribute>
	</xs:complexType>
	</xs:element>
	<xs:element name="Conference">
	<xs:complexType>
		<xs:sequence>
		<xs:element ref="Name"/>
		</xs:sequence>
		<xs:attribute name="semantics" use="required">
		<xs:simpleType>
		<xs:restriction base="xs:string">
			<xs:enumeration value="Information System IS"/>
		</xs:restriction>
		</xs:simpleType>
		</xs:attribute>
		<xs:sequence>
		<xs:element ref="Year"/>
		</xs:sequence>
		<xs:attribute name="semantics" use="required">
		<xs:simpleType>
```

```
                        <xs:restriction base="xs:string">
                                <xs:enumeration value="Null"/>
                        </xs:restriction>
                        </xs:simpleType>
                        </xs:attribute>
                        </xs:complexType>
                </xs:element>
<xs:element name="Abstract">
<xs:attribute name="semantique" use="required">
                <xs:simpleType>
                <xs:restriction base="xs:string">
                <xs:enumeration value="DataBase DB"/>
                </xs:restriction>
                </xs:simpleType>
</xs:attribute>
<xs:attribute name="contenue" use="required">
                <xs:simpleType>
                <xs:restriction base="xs:string">
                <xs:enumeration value="xlink:href= &apos;C:\Doc2\Paper.xml&apos;"/>
                </xs:restriction>
</xs:simpleType>
</xs:attribute>
</xs:schema>
```

Figure 40. XML Schéma du méta-document des structures logique et sémantique.

Dans la section suivante, nous introduisons les langages de requêtes pour les documents XML, et nous décrivons plus particulièrement le langage XQuery qui sera utilisé lors de l'interrogation de l'entrepôt de documents.

4.4. Langages de requêtes pour les documents XML

Pour manipuler les documents XML semi-structurés et plus précisément en extraire des informations en termes de structure et/ou de contenu, de nombreux langages de requêtes ont été proposés. Ces langages devraient prendre en compte non seulement le contenu mais aussi la structure sous-jacente car cette dernière peut changer complètement la pertinence et l'adéquation des documents vis-à-vis des besoins exprimés par la requête.

Les langages de requête de la littérature diffèrent par leur niveau d'expressivité. Nous distinguons deux types de langages de requêtes : Langages textuels et langages visuels (Harrathi R. et Calabretto S., 2010). Les langages textuels peuvent être classés en quatre catégories selon le mode d'utilisation de la structure du document dans les requêtes :

- **Langages textuels basés sur les mots-clés**

Ces langages sont similaires aux langages plein texte utilisés dans la recherche d'information traditionnelle où la requête est exprimée par une conjonction de mots-clés. La caractéristique de ces langages est qu'ils retournent comme résultat un élément (ou nœud) du document XML. XRANK (Guo et al., 2003) et XKSEARCH (Xu et al., 2005) sont des exemples de ces langages textuels basés sur les mots-clés.

- **Langages textuels basés sur les mots-clés et les balises**

Les langages de cette catégorie permettent d'exprimer des contraintes comme le nom de la balise concernée, i.e. qui « annote » le mot-clé. . Nous pouvons citer comme exemple le langage XSEARCH (Cohen et al., 2003).

Dans la requête XSEARCH suivante : "title : Query languages" implique que seul l'élément "title" contenant les mots-clés "Query languages" sera retourné.

- **Langages textuels basés sur les mots-clés et les chemins**

Cette famille de langages permet d'exprimer des contraintes sur la structure en spécifiant des chemins. Des exemples de ces langages sont : XPath 2.0 (XPath 2.0, 2007), XIRQL (Fuhr N. et Großjohann K., 2001) et NEXI (Trotman et al., 2005).

XPath (XML Path Language) permet de sélectionner et de manipuler des parties d'un document XML. Il est essentiellement utilisé pour les deux objectifs suivants :

- Trouver et renvoyer des parties spécifiques d'un document XML. Pour ce faire, XPath convertit d'abord le document en arbre XML. En fonction du chemin d'accès indiqué, XPath peut ensuite trouver et renvoyer l'information demandée.
- Effectuer des opérations sur les parties sélectionnées. Pour cela, il dispose de fonctions prédéfinies lui permettant de réaliser des traitements arithmétiques de base ou de modifier le format des données renvoyées par un chemin d'accès.

- **Langage basé sur le mots-clés et XQuery**

XQuery 1.0 (XML Query Language) sert à sélectionner du contenu dans un document XML, à le transformer et à renvoyer le résultat en format XML. Il hérite des caractéristiques de plusieurs autres langages. De XPath et de XQL, il reprend la syntaxe d'expression de chemin pour l'adressage d'éléments dans les documents XML. De XML-QL, il hérite la notion de variable obligatoire et crée de nouvelles structures. De SQL (Structured Query Language) il reprend l'idée d'une série de clauses basées sur des mots-clés qui fournissent un modèle pour la restructuration des données (SELECT-FROM-WHERE de SQL). Les requêtes basiques de XQuery sont identiques à celles définies par XPath. Si l'on désire faire des requêtes simples, XPath peut donc parfaitement suffire. XQuery est intéressant dès que l'on désire faire des requêtes complexes avec des expressions du type FLWR (For-Let-Where-Return) ou encore faire appel à la récursivité.

XQuery s'appuie sur des fonctions de recherche plein texte ; en particulier un prédicat "*contains* " est intégré pour la recherche par mots-clés.

Afin d'interroger les structures sémantiques, nous avons opté pour le langage de requêtes XQuery pour les raisons suivantes : (1) Possibilité d'exprimer des requêtes complexes, (2) Ressemblance de sa syntaxe à celle de SQL, et (3) Possibilité de formuler une requête sur

plusieurs documents à la fois (ce qui n'est pas le cas pour XSLT[35] (W3C2, 2003) par exemple). Cependant, les langages visuels, caractérisés par l'incorporation et l'interrogation d'éléments non textuels (formulaires, icônes, images, graphes, ...), représentent actuellement la meilleure alternative. Mais vu que dans le processus proposé nous nous focalisons sur l'interrogation sémantique des éléments textuels, nous préférons travailler avec les langages d'interrogation textuels.

Dans la section suivante, nous nous intéressons à l'interrogation en exploitant la structure sémantique. Notre objectif est de montrer la faisabilité de cette interrogation basée sur la structure sémantique que nous avons définie. Nous distinguons deux types de requêtes : les requêtes de type RI et les requêtes de type OLAP.

4.5 Interrogation sémantique dans un contexte RI

Dans ce qui suit, nous proposons un ensemble de requêtes XQuery qui illustrent les modes d'interrogation du contenu des documents en tenant compte de leur structure sémantique décrite dans les méta-documents (tels que définis dans la section 3).

Les exemples de ce chapitre ont été exprimés sur une collection de 20 documents dont un exemple de document est présenté dans Annexe conformes à la structure générique *Paper* de la Figure 32. Ces documents sont numérotés séquentiellement pour faciliter l'établissement du lien entre le document et son code généré automatiquement dans le méta-document correspondant.

Exemple 1 : Trouver les documents dont le titre est associé au concept "*warehouse*".

```
xqueryversion"1.0";
<Résultat>
{
(: NB est le nombre de documents :)
For $j in (1 to NB),   //NB est le nombre de documents
$i in doc(concat ("doc", $j,".xml"))/Paper
Let $titre:=$i/Title
return
    (
if (    fn:contains($title , "warehouse")) then
(
<Document>{concat ("doc", $j,".xml")} </Document>
                )
else
                ()
             )
}</Résultat>
```

[35]XSLT est un langage permettant de produire un document XML ou texte à partir d'un autre document par application de règles de transformation. Par exemple pour afficher un document XML sur un navigateur web, en le convertissant en XHTML, XSLT permet ainsi une mise en forme des données plus propice à l'impression ou à l'affichage sur un terminal d'ordinateur et une conversion du contenu d'un document en un format plus aisément manipulable.

Résultat de l'exemple 1 :

Cette requête retourne deux identifiants de deux documents.

```
<Résultat>
<Document>doc18.xml</Document>
<Document>doc20.xml</Document>
</Résultat>
```

Exemple 2 : Trouver les documents dont le résumé est associé au concept "OLAP".

```
xqueryversion"1.0";
<Résultat>
{
For $j in (1 to 20),
$i in doc(concat ("doc", $j,".xml"))/Paper/Abstract
return
    (
if (    fn:contains(fn:lower-case($i) , "olap") ) then
(
<Document>{concat ("doc", $j,".xml")} </Document>
               )
else
             ()
           )
}</Résultat>
```

Résultat de l'exemple 2 :

```
<Résultat>
<Document>doc1.xml</Document>
<Document>doc2.xml</Document>
<Document>doc3.xml</Document>
<Document>doc6.xml</Document>
<Document>doc8.xml</Document>
<Document>doc9.xml</Document>
<Document>doc14.xml</Document>
<Document>doc15.xml</Document>
<Document>doc16.xml</Document>
<Document>doc18.xml</Document>
<Document>doc20.xml</Document>
</Résultat>
```

Exemple 3 : Trouver les titres des documents qui traitent d'au moins un des deux concepts "OLAP" ou "data warehouse".

```
xqueryversion"1.0";
<Résultat>
{
For $j in (1 to 20),
$k indoc(concat ("doc", $j,".xml"))/Paper,
$iindoc(concat ("méta_document_doc", $j,".xml"))/Balise/@semantique/string()
Let $sem:=$i
return
    (
if(fn:lower-case($i)    =  "olap"or fn:lower-case($i)    =  "data warehouse" ) then
(
<Title>  {$k/Title/string()}              </Title>

)else
      ()

   )

}</Résultat>
```

Résultat de l'exemple 3 :

```
<Résultat>
<Title> Modele multidimensionnel en diamant dedie a l"OLAP
semantique de documents </Title>
<Title> A Novel Multidimensional Model for the OLAP on documents: Modeling,
Generation and Implementation </Title>
<Title> Diamond multidimentional(Dimension and fact...) model and aggregation
operators for document OLAP </Title>
<Title> DWEv: Un prototype pour l'evolution partielle du schema
multidimensionnel</Title>
<Title> Modelisation des transformations pour l'evolution de modeles
multidimensionnels </Title>
<Title> Toward Evolution Models for Data Warehouses</Title>
<Title> Schema multidimensionnel dedie pour l'OLAP des Tweets </Title>
<Title> OLAP of the tweets: From modeling toward exploitation </Title>
<Title> Toward Propagating the Evolution of Data Warehouse on Data Marts</Title>
<Title> A new multidimensional model for the OLAP of documents based on facets
</Title><Title> Building an XML document warehouse</Title>
</Résultat>
```

Exemple 4 : Trouver les concepts qui apparaissent plus d'une fois dans le document (dans le Titre et/ou dans le Résume).

```
xquery version"1.0";
<Concepts>
{
for $i in (1 to 20),
$b in doc(concat ("doc", $i,".xml"))/Paper
for $p in distinct-values( collection("*")/Balise/@semantique)
for $sem in $p
return
  (
let $toks_caption:= $b /Abstract/text()/fn:tokenize(fn:normalize-space(.),'\s'),
$toks_title:= $b /Title/text()/fn:tokenize(fn:normalize-space(.),'\s')
for $t in distinct-values($toks_caption)
let $count:= count($toks_caption[. = $t])
let $count1:= count($toks_title[. = $t])
return
            (
if(  $count>1  and fn:contains( fn:lower-case($sem), fn:lower-case($t) ) and
fn:string-length($t)>=4  ) then
                  (
<Apparition>
<Document>       {concat ("doc", $i,".xml")  } </Document>
<Concept> {fn:lower-case($sem) }    </Concept>
<Nombre_apparition> {$count}     </Nombre_apparition>
</Apparition>
)else
                                 ()
          )
)
}</Concepts>
```

Résultat de l'exemple 4 :

```
   <Concepts>
      <Apparition>
              <Document> doc2.xml</Document>
              <Concept>olap</Concept>
              <Nombre d'apparition> 2</Nombre d'apparition>
       </Apparition>
       <Apparition>
              <Document>     Doc7.xml    <Document>
              <Concept>    data warehouse      </Concept>
```

```
            <Nombre d'apparition>   3      <Nombre d'apparition>
      </Apparition>
      <Apparition>
            <Document>     Doc20.xml   <Document>
            <Concept>    database </Concept>
            <Nombre d'apparition>   2      <Nombre d'apparition>
      </Apparition>
      <Apparition>
            <Document>     Doc9.xml   <Document>
            <Concept>    information system </Concept>
            <Nombre d'apparition>   2      <Nombre d'apparition>
      </Apparition>
      <Apparition>
            <Document>     Doc19.xml   <Document>
            <Concept>    document warehouse </Concept>
            <Nombre d'apparition>   2      <Nombre d'apparition>
      </Apparition>
  <Concepts>
```

Exemple 5 : Afficher pour chaque concept le nombre de documents associés triés par ordre décroissant du nombre de documents.

```
Xqueryversion"3.0";
<Résultat>
{
For $i in (1 to 20),
$k in doc(concat ("doc", $i,".xml")),
$b in doc(concat ("méta_document_doc", $i, ".xml"))/Balise/@semantique
Let $m:=$b
Orderby count ($k)  descending
Groupby $b
return
 (

<Apparition>
<Concept>   {distinct-values($m/string())}       </Concept>
<Nombre_document>    {count($k) }                          </Nombre_document>

</Apparition>
 )
}
   </Résultat>
```

Résultat de l'exemple 5 :

Ce résultat peut être visualisé par l'utilisateur de la manière suivante :

Concept	Nombre de documents
DATA WAREHOUSE	6
OLAP	5
DESIGN	4
DOCUMENT WAREHOUSE	2
INFORMATION SYSTEM	2
Data Base	1

4.6 Interrogation OLAP de la structure sémantique

Dans le domaine des bases de données multidimensionnelles, le traitement analytique en ligne (OLAP :"*On-Line Analytical Processing*") est une technologie orientée vers l'analyse instantanée d'informations selon plusieurs axes, dans le but de fournir des rapports de synthèse tels que ceux utilisés en analyse financière. Les applications de type OLAP sont

couramment utilisées en informatique décisionnelle, dans le but d'aider les décideurs à construire une vue transversale de l'activité d'une entreprise (Kimball, 2002).

L'OLAP consiste à considérer un sujet analysé comme un point dans un espace à plusieurs dimensions. Les données sont organisées de manière à mettre en évidence le sujet analysé (appelé fait) et les différentes perspectives ou axes d'analyse (dimensions). Ces dimensions possèdent des hiérarchies permettant de réaliser des analyses à différents niveaux de granularité (niveaux de détail).

Exemple d'OLAP sur le contenu sémantique

Dans notre cas, le point de départ de cette requête OLAP est une collection de documents XML de structures identiques. Supposons que nous souhaitons analyser le nombre de documents selon les dimensions *Author*, *Year* et selon le *Topic*. Dans cet exemple, la thématique traite l'aspect sémantique des documents. Pour réaliser cette requête OLAP, nous devons développer 5 requêtes intermédiaires qui sont :

1. Déterminer l'auteur de chaque document.
2. Trouver l'année de chaque document.
3. Retrouver le concept associé à l'élément *Abstract* de chaque document.
4. Relier les résultats des trois requêtes précédentes par jointure (sur le nom du document).
5. Grouper par *Author*, *Year* et *Topic* et calculer le nombre de documents.

Les requêtes 1 à 5 suivantes, exprimées en XQuery, traduisent les 5 requêtes intermédiaires décrites précédemment. Un extrait du résultat de chaque requête est également donné.

Requête 1 : Déterminer l'auteur "Author" de chaque document.

```
xqueryversion"1.0";
<Documents>
    {
For $j in (1 to 20),
$i in doc(concat ("doc", $j,".xml"))/Paper
Let $aut:=$i/Author
return
   (
<Document> {concat ("doc", $j,".xml") }
<Author> { $aut/text() }    </Author>
</Document>
    )
}</Documents>
```

Résultat de la requête 1 :

```
<Documents>
<Document>doc1.xml
<Author> Salma Ben Meftah </Author>
</Document>
<Document>doc2.xml
<Author> Salma Ben Meftah </Author>
</Document>
```

```
   <Document>doc3.xml
   <Author> Salma Ben Meftah </Author>
   </Document>
   <Document>doc4.xml
   <Author> Kais Khrouf </Author>
   </Document>
   <Document>doc5.xml
   <Author> Kais Khrouf </Author>
   </Document>
   …
   </Documents>
```

Requête 2 : Trouver l'année "Year" de chaque document.

```
xqueryversion"1.0";
<Documents>
    {
for$jin (1 to 20),
$iindoc(concat ("doc", $j,".xml"))/Paper/Conference/Year
Let $annee:=$i
return
   (
<Document> {concat ("doc", $j,".xml") }

<Year>{ $annee/text() }    </Year>
</Document>
    )
}</Documents>
```

Résultat de la requête 2 :

```
<Documents>
<Document>doc1.xml
<Year>  2013  </Year>
</Document>
<Document>doc2.xml
<Year>  2013  </Year>
</Document>
<Document>doc3.xml
<Year>  2014  </Year>
</Document>
<Document>doc4.xml
<Year>  2013  </Year>
</Document>
<Document>doc5.xml
<Year>  2013  </Year>
</Document>
…
</Documents>
```

Requête 3 : Retrouver le concept associé à l'élément *Résumé* "Abstract" de chaque document.

```
xqueryversion"1.0";
<Documents>
  {
for$iin (1 to 20),
$aindoc(concat ("méta_document_doc", $i, ".xml"))/Balise/Balise
return
  (
if ($a/@Nom ="Abstract" ) then
(
<Document>
{ concat ("doc", $i,".xml") }
<Semantique>   {$a/@semantique} </Semantique>
</Document>
    ) else
```

```
        ()
    )

    }  </Documents>
```

Résultat de la requête 3 :

```
    <Documents>
    <Document>doc1.xml
    <Semantique semantique="Indexation Sémantique"/>
    </Document>
    <Document>doc2.xml
    <Semantique semantique="Indexation Sémantique"/>
    </Document>
    <Document>doc3.xml
    <Semantique semantique="Indexation Sémantique"/>
    </Document>
    <Document>doc4.xml
    <Semantique semantique="OLAP"/>
    </Document>
    <Document>doc5.xml
    <Semantique semantique="OLAP"/>
    </Document>
    …
    </Documents>
```

Requête 4 : Relier les résultats des trois requêtes précédentes par jointure.

```
xqueryversion"1.0";
<Documents>
{
For $j in (1 to 14)
For $auteur in doc("document_auteur.xml")/Documents
return
      (
For $annee in doc("document_Annee.xml")/Documents
return
(
For $concept in doc("document_Concept.xml")/Documents
return
(
if($auteur/Document/text()= $annee/Document/text() and$auteur/Document/text() =
$concept/Document/text() ) then
            (
<Document> {$auteur/Document[$j]/text()}
<Author> {$auteur/Document[$j]/Author/text()}  </Author>
<Year>   {   $annee/Document[$j]/Year/text()    } </Year>
<Concept>{$concept/Document[$j]/Semantique/@semantique} </Concept>
</Document>
)else
                  ()
        )
     )
)
} </Documents>
```

Résultat de la requête 4 :

```
<Documents>
<Document>doc1.xml
<Author> Salma Ben Meftah </Author>
<Annee>  2013  </Annee>
<Concept semantique="Indexation Sémantique"/>
</Document>
<Document>doc2.xml
<Author> Salma Ben Meftah </Author>
```

```
<Annee>  2013  </Annee>
<Concept semantique="Indexation Sémantique"/>
</Document>
<Document>doc3.xml
<Author> Salma Ben Meftah </Author>
<Annee>  2014  </Annee>
<Concept semantique="Indexation Sémantique"/>
</Document>
<Document>doc4.xml
<Author> Kais Khrouf </AuthorAuthor>
<Annee>  2013  </Annee>
<Concept semantique="OLAP"/>
</Document>
<Document>doc5.xml
<Author> Kais Khrouf </Author>
<Annee>  2013  </Annee>
<Concept semantique="OLAP"/>
</Document>
…
</Documents>
```

Requête 5 : Regrouper par *Auteur"Author"*, *Année "Year"* et *Thématique "Topic"* et calculer le nombre de documents de chaque partition.

```
xqueryversion"3.0";
<Cube>
  {
for$jindoc("jointure_auteur_annee_concept.xml")/Documents/Document
groupby$auteur:=$j/Year, $année:=$j/Year, $concept:= $j/Concept/@semantique
return
   (
<Cellule>
<Author> {$auteur} </Author>
<Year> {$année}     </Year>
<Concept> {$concept}     </Concept>
<Nombre>{count($j)} </Nombre>

</Cellule>
    )
}</Cube>
```

Résultat de la requête 5 :

Le résultat peut être schématisé à l'aide d'un tableau multidimensionnel comme le montre le tableau ci dessous :

Count (Nombre de documents)		Topic		...
		Indexation sémantique	OLAP	...
Author	Year			
Salma Ben Meftah	2013	2		
	2014	1		
Kais Khrouf	2013	2		
	2013		3	
...	...	...	...	...

En utilisant XQuery, les utilisateurs peuvent interroger les documents XML en se basant uniquement sur leurs structures logiques et leurs contenus. Les méta-documents, que nous proposons, permettent ainsi de représenter la sémantique véhiculée par ces documents XML ; ce qui permet aussi aux utilisateurs de profiter de la structure sémantique et d'en déduire de nouvelles informations et connaissances.

4.7. Conclusion

Nous avons consacré ce chapitre à montrer l'utilité de la structure sémantique, construite dans le chapitre 3, dans un processus d'interrogation. Pour ce faire, nous avons commencé par étendre le méta-modèle d'entrepôts de documents en lui intégrant la structure sémantique. Plus particulièrement, nous avons ajouté dans ce méta-modèle une partie sémantique contenant deux classes : *Taxonomies* et *Concepts* (cf. Figure 35).

Nous avons aussi proposé deux solutions pour le stockage des structures sémantiques et nous avons retenu celle qui repose sur la définition de méta-documents. Par la suite, nous avons présenté l'interrogation sémantique à travers un ensemble de requêtes dans un contexte de RI, et des requêtes dans un contexte OLAP, en utilisant le langage de requête XQuery.

L'objet de ces requêtes est de montrer la faisabilité de l'interrogation sémantique des documents XML. Néanmoins, il serait intéressant de proposer à l'utilisateur soit un langage spécifique voire une interface graphique pour exprimer son besoin d'une manière beaucoup plus simple. Par la suite, c'est le système qui se charge de la génération du script correspondant en XQuery.

Le chapitre suivant est dédié aux expérimentations qui nous permettent d'évaluer et de valider nos propositions.

Chapitre 5

Expérimentations et évaluation

Sommaire

5.1 Introduction

Comme nous l'avons mentionné dans le chapitre 3, l'objectif de l'approche proposée dans le cadre de cette thèse est la structuration sémantique des documents XML à partir de leur structure logique spécifique, de leur contenu et, par référence à une taxonomie de domaine sélectionnée. Pour valider notre approche et afin de vérifier l'utilité et montrer l'apport de la pondération des taxonomies dans la sélection, nous menons dans ce chapitre une étude expérimentale de nos propositions. Pour ce faire, nous utilisons : 1) Le thésaurus *MeSH* spécialisé dans le domaine médical, à partir duquel nous avons constitué plusieurs taxonomies de domaines. Nous avons choisi ce thésaurus à cause de la richesse de sa structure hiérarchique ; et 2) La collection de documents XML*ImageCLEFMed 2010* comme base de test.

Dans ce chapitre, nous commençons par introduire l'environnement et les outils de développement utilisés. Ensuite, nous présentons la base de test utilisée dans notre expérimentation (XML*ImageCLEFMed 2010*) et le thésaurus *MeSH*. Nous détaillons par la suite l'architecture du prototype développé et nous décrivons chacune de ses composantes. Nous terminons ce chapitre par une présentation des expérimentations réalisées et des résultats obtenus.

5.2 Outils utilisés

Afin de réaliser nos expérimentations, nous avons utilisé :

- le langage de programmation Java, avec Eclipse comme environnement de développement ;
- JDOM "Java Document Object Model" pour la manipulation des documents XML en Java ;
- le Système de Gestion de Bases de Données (SGBD) Oracle 9i pour le stockage et l'accès aux données ;
- ETL OPEN TALEND comme processus d'intégration et de gestion de grands volumes de données.

JAVA

Pour le développement de notre prototype, nous avons opté pour le langage Java et cela pour de nombreuses raisons :

- Java est un langage orienté objet, il est caractérisé aussi par la réutilisation de son code ainsi que la simplicité de sa mise en œuvre ;

- Il est indépendant de toute plateforme, il est possible d'exécuter des programmes Java sur tous les environnements qui possèdent une JVM "*Java Virtual Machine*" ;
- Il est doté d'une riche bibliothèque de classes, comprenant la gestion des interfaces graphiques (fenêtres, menus, graphismes, boîtes de dialogue, contrôles) et la gestion des exceptions ;
- Il permet d'accéder d'une manière simple aux fichiers et aux réseaux (notamment Internet) ;
- Il permet un accès simplifié aux bases de données, soit à travers la passerelle JDBC-ODBC "Java-*Object Database Connectivity*" ou à travers un pilote JDBC "*Java Database Connectivity*" spécifique au SGBD ;
- Il permet de manipuler des données XML à travers des API "*Application Programming Interface*" XML de Sun.

JDOM

JDOM est une API du langage Java. Elle permet de modéliser, de parcourir et de manipuler un document XML.

L'API de JDOM propose les fonctionnalités suivantes :

- Lecture de fichiers XML à partir de fichiers, arbres DOM *"Document Object Model*", flux SAX *"Simple API for XML*" ;
- Création de documents XML ;
- Exportation d'arbre XML JDOM sous la forme de fichier, arbre DOM, flux SAX.

A ce niveau, la question qui se pose est : *qu'est-ce que nous apporte de plus le JDOM ?* La réponse est sa simplicité grâce à l'utilisation de la technique des filtres qui nous permet de travailler sur un sous-ensemble d'éléments (granules) de documents répondant à un certain nombre de critères.

Accès à la base de données Oracle 9i avec Java :

Le paquetage java.sql d'Eclipse contient des classes qui fournissent l'API permettant d'accéder à une source de données. Il est également appelé API JDBC 2.0. Le paquetage java.sql contient des classes, interfaces et méthodes permettant d'établir des connexions aux bases de données, d'envoyer des instructions, de récupérer le résultat d'une requête...

D'autres API ont été construites au-dessus de JDBC telles que l'API SQLJ (cf. Figure 42). Elle est composée de 3 composantes principales (Sophia A. & Richard G., 2006) :

- La première partie de SQLJ[36] (fichier source .sqlj) fait partie de SQL3 : elle définit des extensions de SQL pour permettre d'insérer directement des ordres SQL au milieu d'un programme Java ;
- Un pré-compilateur transforme les ordres SQL en des instructions Java contenant des appels de méthodes JDBC. Le résultat est un programme source Java que l'on peut compiler avec un compilateur Java ordinaire. Ce pré-compilateur contrôle la validité des ordres SQL ;
- Un compilateur Java (exemple javacc) : Outil qui permet de lire une spécification grammaticale et de la convertir en un programme Java capable de reconnaître les correspondances de cette grammaire.

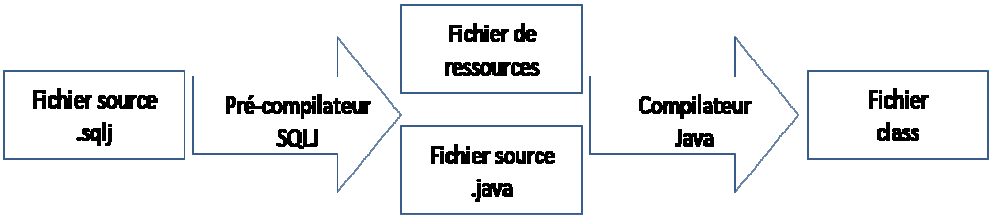


Figure 42 : Principe de SQLJ (Sophia A. & Richard G., 2006).

ETL OPEN TALEND

ETL signifie Extract, Transform, Load (Extraction, Transformation, Chargement). Ce sont les trois étapes que doit impérativement réaliser un outil ETL.

Nous avons utilisé l'outil OPEN TALEND afin de charger des données à partir de tous les domaines de thésaurus *MeSH* et de toute la collection *XMLImageCLEFMed 2010*. Vu la grande quantité de données stockées au format XML dans *MeSH* et dans la collection *XMLImageCLEFMed 2010,* nous avons besoin d'un *ETL* afin d'extraire les données à partir d'un document XML et de les charger dans une base de données relationnelles (entrepôt de documents).

L'avantage de cet *ETL* réside dans sa rapidité à charger de grands volumes de données.

5.3 Base de test.

5.3.1 Collection de documents XML *ImageCLEFMed* 2010

L'ensemble des données de la collection *XMLImageCLEFMed 2010* est fournie par *RSNA*[37] "*Radiological Society of North America*". Cette base compte 77 506 images décrites par les informations suivantes :

- <figureID> :'identifiant de l'image,

[36] *SQLJ est un standard ANSI développé par un consortium des plusieurs éditeurs de logiciels pour serveurs d'application et de bases de données, tels que IBM® Corporation, Microsoft Corporation, Sun Microsystems et Oracle. Le traducteur SQLJ convertit un fichier source SQLJ en un fichier source Java*

[37] *http://www.rsna.org/*

- <figureURL> : URL de l'image,
- <caption> : légende de l'image,
- <title> : titre de l'article à partir duquel l'image a été extraite,
- <articleURL> : URL de l'article contenant l'image.

La Figure 43 est un extrait de la collection *XMLImageCLEFMed 2010*.

<imageclef>

<record>

<figureID>27981**</figureID>**

<**figureURL**>http://radiology.rsnajnls.org/cgi/content/full/210/1/11/F3</**figureURL**><**caption**><B>Figure3.</B><B></B> Photograph illustrates the technique of thymic irradiation for either status thymicolymphaticus or thymic asthma. In the caption for the original illustration, note was made that a piece of lead rubber sheeting normally covered the restrained infant, except for a window over the thymus. In this photograph, the sheeting was not present to demonstrate the "relation of the cone to the chest." (Reprinted, with permission, from reference 32.)<P>
</**caption**>

<**title**>The right place at the wrong time: historical perspective of the relation of the thymus gland and pediatric radiology</**title**>

<pmid>9885579</pmid><**articleURL**>http://radiology.rsnajnls.org/cgi/content/full/210/1/11**</articleURL**>

<**imageLocalName**>27981.jpg**<**/imageLocalName>

</record>

</**imageclef**>

Figure 43. Exemple de document XML décrivant une image, extrait de la collection *XMLImageCLEFMed 2010*.

5.3.2 Thésaurus *MeSH*

Le thésaurus *MeSH* a été initialement créé en anglais. Il est régulièrement mis à jour. La traduction de *MeSH* vers le français est assurée par l'INSERM[38] qui met la version bilingue à la disposition de la communauté francophone.

MeSH comprend essentiellement des termes qui désignent les concepts biomédicaux, des descripteurs, des relations et des qualificatifs. Nous détaillons dans ce qui suit ces éléments et leurs rôles :

- **Terme :** un terme est un mot ou un ensemble de mots exprimant une notion particulière ;
- **Concept :** un concept comprend un ou plusieurs termes synonymes et porte le nom d'un de ces termes, dit « *terme préféré* » ;

[38]http://www.inserm.fr/

– **Relation :** dans *MeSH*, il existe deux types de relations entre concepts : les *relations hiérarchiques* et les *relations associatives* (associé à) comme suit :

1. La relation « *est une partie de* » (méronymie). Par exemple, le concept « Pain » (C10.597.617) est une partie de « Neurologic Manifestations » (C10.597).
2. La relation « *est un type de* » (hyponymie). Par exemple, le concept « Neurologic Manifestations » (C10.597) est un type de « Nervous System Diseases » (C10).
3. La relation « *est sémantiquement proche de* » (homonymes). Par exemple, le concept « Acute Pain » (C10.597.617.088) est sémantiquement proche de « Neurologic Manifestations » (C10.597).

La hiérarchisation entre concepts est représentée par un code (par exemple C10 dans la figure 44 pour le concept « Nervous System Diseases ») identifiant l'arborescence auquel le concept appartient.

La Figure 44 illustre un exemple d'arborescence extrait de *MeSH* (le symbole + permet d'étendre l'affichage des sous concepts).

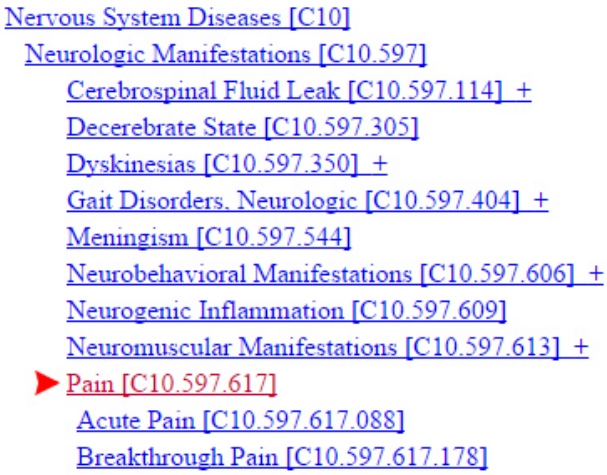


Figure 44. Extrait de l'arborescence *C10* (domaine "*Diseases*") de *MeSH*.

– **Descripteur** : Un descripteur est constitué d'un ou de plusieurs concepts de significations proches et porte le nom d'un de ces concepts, dit préféré. Les autres concepts, dits subordonnés, présentent une relation sémantique avec le concept préféré, soit une relation hiérarchique (générique ou spécifique), soit une relation associative (associé).

Les descripteurs *MeSH* sont répartis en 16 domaines[39] recouvrant la biologie, la médecine et les domaines connexes (cf. Tableau 5).

	Nom Domaine *MeSH*
[A]	Anatomy
[B]	Organisme
[C]	Maladies

[39] http://www.nlm.nih.gov/cgi/mesh/2014/MB_cgi

[D]	Produits chimiques et pharmaceutiques
[E]	Techniques analytiques, diagnostiques, /thérapeutiques et équipements
[F]	Psychiatrie et psychologie
[G]	Phénomènes et processus
[H]	Disciplines et professions
[I]	Anthropologie, enseignement, sociologie, et phénomènes sociaux
[J]	Technologie, industrie et agriculture
[K]	Sciences humaines
[L]	Sciences de l'information
[M]	Individus
[N]	Santé
[V]	Caractéristiques d'une publication
[Z]	Lieux géographiques

Tableau 5. Les seize domaines de *MeSH*.

Chaque domaine de descripteurs est structuré selon une arborescence (hiérarchie de concepts) pouvant comprendre jusqu'à onze niveaux de hiérarchies. Chaque descripteur est représenté par un code alphanumérique, la lettre indiquant le domaine et la séquence numérique précisant la localisation (niveau) dans la hiérarchie. Certains descripteurs ont plusieurs localisations, au sein du même domaine ou de domaines différents, et plusieurs codes alphanumériques représentant chacun une localisation (Tree_number en *MeSH*). Par exemple, le descripteur « Pain » appartient à plusieurs hiérarchies (cf. Figure 45) : C10.597.617, C23.888.592.612, C23.888.646 et F02.830.816.444.

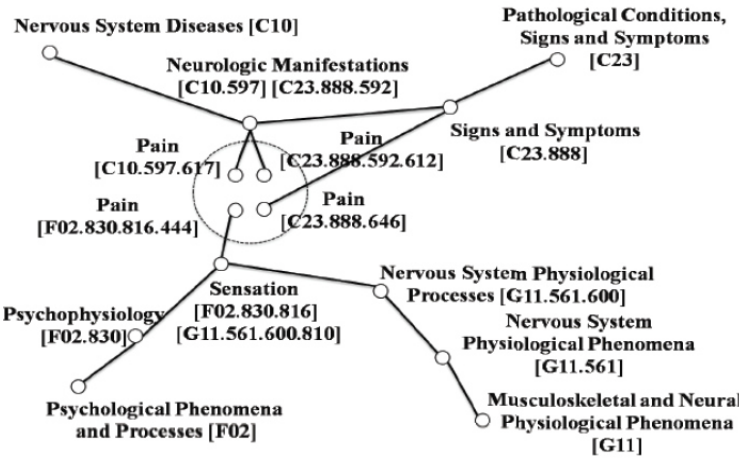


Figure 45. Descripteur «Pain» appartenant à plusieurs hiérarchies dans *MeSH*.

Remarque : Dans nos travaux et afin de vérifier et valider la première phase de notre approche (choix de taxonomie), nous avons considéré que chaque descripteur du thésaurus *MeSH* du premier niveau est une taxonomie à part. De cette façon, nous obtenons 128 taxonomies de thématiques différentes.

Pour bien et mieux comprendre les éléments de MeSH, il est important de savoir que :

1) La structure des données de *MeSH* est à trois niveaux (les descripteurs, les concepts, les termes)

2) La référence à ces objets se fait à la fois par un nom unique et un identifiant unique (Descriptor UI, Terme UI)

3) La représentation de la structure des éléments d'occurrence multiple est effectuée grâce à l'utilisation de listes "*TreeNumberList, ConceptList, TermList*".

Nous présentons dans ce qui suit la liste des éléments disponibles pour les données MeSH en format XML pour le descripteur "Abortifacient Agents" (cf. Figure 46) y compris les hiérarchies (balise TreeNumberList), les concepts (balise ConceptList) et les termes (balise TermList).

```
// descripteur possèdant comme identifiant : D000019
 <DescriptorU >D000019</DescriptorUI>
   <DescriptorName>
    <String> Abortifacient Agents </String>
  </DescriptorName>

// l'ensemble des hierarchies où se localise le descripteur "Abortifacient Agents " possédant comme identifiant D00019

        <TreeNumberList>
          <TreeNumber>D27.505.696.875.131</TreeNumber>
          <TreeNumber>D27.505.954.705.131</TreeNumber>
        </TreeNumberList>

// Liste des concepts pour le descripteur "Abortifacient Agents " : le concepts préféré
 <ConceptList>
    <Concept PreferredConceptYN="Y">
           <RegistryNumber>0</RegistryNumber>
           <ScopeNote>Chemical substances that interrupt pregnancy after implantation.</ScopeNote>
           <TermList>

                  <Term ConceptPreferredTermYN="N" IsPermutedTermYN="Y" LexicalTag="NON" PrintFlagYN="N" RecordPreferredTermYN="N">
                   <TermUI>T000046</TermUI>
                    <String>Agents, Abortifacient</String>
                  </Term>

                  <Term ConceptPreferredTermYN="N" IsPermutedTermYN="N" LexicalTag="NON" PrintFlagYN="N" RecordPreferredTermYN="N">
                   <TermUI>T000048</TermUI>
                   <String>Abortifacients</String>
                  </Term>

                  <Term ConceptPreferredTermYN="N" IsPermutedTermYN="N" LexicalTag="NON" PrintFlagYN="N" RecordPreferredTermYN="N">
                   <TermUI>T000047</TermUI>
                   <String>Contraceptive Agents, Postconception</String>
                  </Term>

                  <Term ConceptPreferredTermYN="N" IsPermutedTermYN="Y" LexicalTag="NON" PrintFlagYN="N" RecordPreferredTermYN="N">
                   <TermUI>T000047</TermUI>
                   <String>Agents, Postconception Contraceptive</String>
                  </Term>

                  <Term ConceptPreferredTermYN="N" IsPermutedTermYN="Y" LexicalTag="NON" PrintFlagYN="N" RecordPreferredTermYN="N">
                   <TermUI>T000047</TermUI>
                   <String>Postconception Contraceptive Agents</String>
                  </Term>
        </TermList>
     </Concept>
     ...
 </ConceptList>
```

Figure 46. Extrait du thésaurus *MeSH* pour le descripteur "*Abortifacient Agents*" au format *XML*.

5.4 Approche de structuration sémantique

La Figure 47 montre les étapes générales du prototype que nous avons développé pour l'expérimentation et la validation de notre approche de structuration sémantique de documents XML.

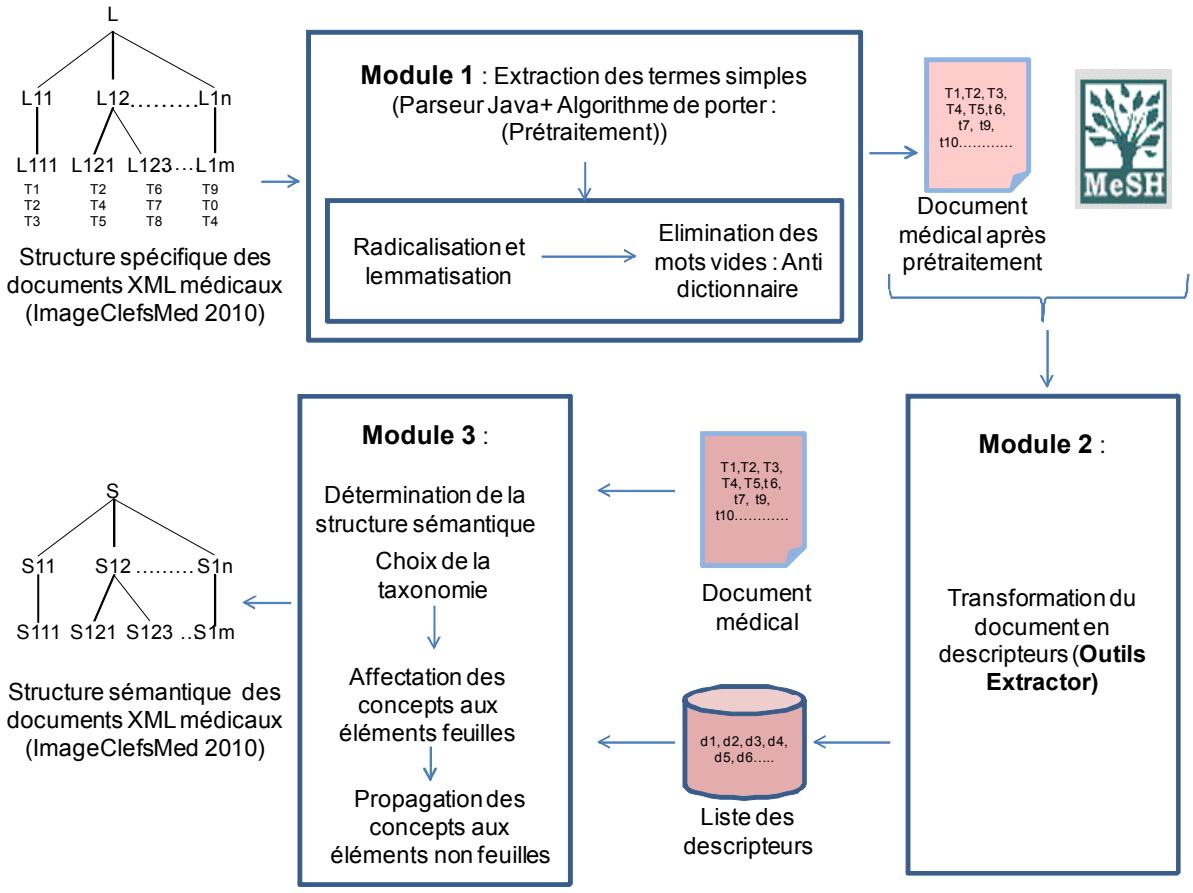


Figure 47. Étapes d'expérimentation de notre approche de structuration sémantique.

Ces 3 modules sont détaillés dans les paragraphes qui suivent.

5.4.1 Extraction des termes simples (Module 1)

L'extraction de termes vise à recenser tous les termes contenus dans un corpus. Le résultat d'une extraction est une liste de termes candidats.

Afin d'extraire les termes simples, nous avons programmé un parseur java à base de l'API JDOM. Ces termes seront ensuite lemmatisés (radicalisation). Nous présentons dans la section suivante l'algorithme de Porter qui est le plus utilisé pour ce type de traitement sur des textes en langue anglaise.

5.4.1.1 Algorithme de Porter

L'algorithme de Porter pour la langue anglaise (cf. chapitre 1, section 1.4.3) se compose d'une cinquantaine de règles de radicalisation/lemmatisation classées en sept phases successives (traitement des pluriels et verbes à la troisième personne du singulier, traitement

du passé et du progressif, etc.). Les mots à analyser passent par tous les stades et, dans le cas où plusieurs règles pourraient leur être appliquées, c'est toujours celle comprenant le suffixe le plus long qui est choisie. La radicalisation/lemmatisation est accompagnée, dans la même étape, de règles de recodage. Ainsi, par exemple, "troubling" deviendra "troubl" par enlèvement du suffixe marqueur du progressif -ing et sera ensuite transformé en "trouble" par application de la règle *"bl" devient "ble"*. Cet algorithme comprend aussi cinq règles de contexte, qui indiquent les conditions dans lesquelles un suffixe devra être supprimé. La terminaison en -ing, par exemple, ne sera enlevée que si le radical comporte au moins une voyelle. De cette manière, "sing" restera "sing".

5.4.1.2 Elimination des mots vides : Anti-dictionnaire

Cette étape consiste à éliminer les mots vides en utilisant un anti-dictionnaire contenant une liste de tous les « stop-words » (mots vides) qui sont généralement (the, on, of, for, in ...).

Exemple : après application du module 1 de prétraitement, la phrase « *La prise en charge du diabète de type 2* », sera transformée comme suit : « prise charge diabète type ».

5.4.1.3 Exemple pour le Module1 : Prétraitement

La figure 48 montre les résultats des principales étapes d'extraction des termes simples (prétraitement).

Photograph illustrates the technique of thymic irradiation for either status thymicolymphaticus or thymic asthma. In the caption for the original illustration, note was made that a piece of lead rubber sheeting normally covered the restrained infant, except for a window over the thymus.

(a) Document

Photograph illustrate the technique of thymic irradiation for either status hymicolymphaticu or thymic asthma.
In the caption for the original illustrator note make that a piece of lead rubber sheet normal cover the restrain infant except for a window over the thymu.

(b) Lemmatisation du texte (Radicalisation)

Photograph illustrate technique thymic irradiation either status thymic olymphaticu thymic asthma.
caption original illustrator piece lead rubber sheet normal cover restrain infant except window thymu.

(c) Elimination des mots vides

Figure 48. Étape Prétraitement d'extraction des termes simples.

5.4.2 Transformation du document en descripteurs (Module 2)

Dans ce module, nous avons transformé le document prétraité en descripteurs afin d'associer de la sémantique au document. Pour ce faire, nous avons opté et utilisé l'outil *MaxMatcher* présenté au chapitre 1 (cf. Section 1.5.4.4) dans le but de faire la correspondance entre les termes de la collection XMLImageCLEFMed 2010 et les descripteurs de MeSH.

Les résultats récupérés par *MaxMatcher* sont constitués comme suit (cf. Figure 49) : Chaque terme identifié est mis en gras, ses concepts associés sont mis entre parenthèses avec un identifiant unique, et sa forme préférée est en italique. La liste des concepts identifiés est également présentée (cf. Figure 49, deuxième partie). Chaque ligne correspond à un concept caractérisé par son rang, son identifiant unique, sa forme préférée, son poids et les hiérarchies dans lesquelles il se trouve.

Exemple : Le concept nommée "*Rubber*" appartient à la hiérarchie possédant comme identifiant (tree number) "C11.768.740".

Partie 1 : Document après prétraitement

<caption> Photograph illustrate technique thymic irradiation either status thymic olymphaticu thymic **asthma**(C0004096, Asthma)**.**
caption original illustrator piece lead rubber sheet normal cover restrain **infant** (*C0021270, Infant*) except window **thymu**(C0040113, *Thymus Gland*). </caption>
<title> right place wrong **time** (C0040223, *Time*) historical perspective relation **thymu gland** (C0040113, *Thymus Gland*) pediatric **radiology** (C0034599, *Radiology*) </title>

Partie 2 : Concepts extrait par extractor

0|C0040113|Thymus Gland|27,4131|
3|C0021270|Infant|0,0000|
4|C0035918|Rubber|0,0000|C11.768.740
5|C0004096|Asthma|0,0000|D08.811.682.65

<DOCNO> 2
0|C0034599|Radiology|0,0000|
1|C0040113|Thymus Gland|0,0000|
2|C0040223|Time|0,0000|

Figure 49. Concepts extraits du document de la Figure 43 avec *MaxMatcher*.

5.4.3 Détermination de la structure sémantique (Module 3)

Ce module consiste à déterminer la structure sémantique à partir de la structure logique d'un document XML. Pour ce faire, nous devons déterminer la taxonomie la plus appropriée pour le document XML.

Ensuite, il faut affecter des concepts aux éléments feuilles de la structure logique du document (au moins aux éléments de type texte en suivant la démarche présentée dans le Chapitre 2.

Finalement, nous propageons les concepts vers les éléments non-feuilles de la structure sémantique en cours de construction (inférence des concepts) et ce en utilisant les 3 règles définies (cf. Chapitre 2).

5.5 Validation et expérimentations

Pour valider nos propositions, nous testons notre approche de structuration sémantique sur un échantillon de la collection de documents *XMLImageCLEFMed 2010* et des taxonomies de domaine issues du thésaurus *MeSH*.

Nous rappelons que les descripteurs de niveau 1 sont considérés dans nos expérimentations comme la racine de taxonomies différentes pour les deux raisons suivantes :

1) Chacun de ces descripteurs décrit un contexte particulier et est décrit par un ensemble de concepts.

2) cette supposition permet de tester et de valider la 1ère phase de notre approche, i.e.Choix d'une taxonomie par document.

Les caractéristiques de la base utilisée pour l'évaluation sont décrites dans le Tableau 6.

Description	Nombre
Documents	41 663
Eléments feuilles	179 340
Eléments non feuilles	51 240
Taxonomies	128
Concepts	25 186
Nombre min / max de niveaux	4 /9

Tableau 6. Caractéristiques de la base de tests.

Afin de vérifier et de valider l'apport de la pondération des taxonomies, nous avons réalisé deux séries de tests : 1) Sans tenir compte des poids des concepts (Algorithme *Sans_Poids*) et 2) En tenant compte de la pondération automatique des concepts des taxonomies (Algorithme *Avec_Poids*). Dans ce qui suit, nous présentons les résultats obtenus dans les deux expérimentations.

Le tableau 7 présente le nombre de taxonomies assignées aux documents selon les deux algorithmes pour la base de test comportant 41 663 documents.

Description	Algorithme Sans_Poids	Algorithme Avec_Poids	Amélioration due à l'algorithme Avec_*Poids*
Nombre de documents ayant une seule taxonomie assignée	11137	25606	34.72 %
Nombre de documents ayant 2 ou plusieurs taxonomies assignées	14482	13	34.72 %
Nombre de documents n'ayant aucune taxonomie assignée	16044	16044	0 %

Tableau 7. Nombre de taxonomies affectées aux documents amélioré par l'algorithme *Avec_Poids*.

Nous remarquons qu'avec l'algorithme *Avec_Poids*, une seule taxonomie a été associée à 25606 documents. Alors que, avec l'algorithme *Sans_Poids*, deux ou plusieurs taxonomies ont

été attribuées à 14482 documents ; ce qui représente une amélioration de 34.72 % obtenue pour l'algorithme Avec_Poids. Aussi, nous n'avons que 13 documents ayant été associés à plus d'une taxonomie pour l'algorithme *Avec_Poids*.

Afin de montrer la pertinence de l'affectation des taxonomies appropriées aux documents, nous avons vérifié manuellement cette affectation sur 1000 documents (ceux qui ont obtenu plus de concepts) de la collection XMLImageCLEFMed 2010.

Dans le tableau 8, nous avons examiné l'association des taxonomies aux 1000 documents de notre échantillon pour savoir celles qui ont été correctement associées.

Description	***Algo. Sans_Poids***	***Algo. Avec_Poids***
Taxonomies correctement associées	875	834
Taxonomies non correctement associées	927	1

Tableau 8. Association des taxonomies aux 1000 documents.

Nous remarquons que l'algorithme *Sans_Poids* affecte plus de taxonomies pertinentes (correctement associées) que l'algorithme *Avec_Poids* ; cela peut s'expliquer par le fait que certains documents abordent deux domaines à la fois (voire plus) et que l'algorithme *Avec_Poids* affecte généralement la taxonomie la plus pertinente par document. Cependant, l'algorithme *Avec_poids* permet de limiter d'une manière très significative les erreurs d'association et d'écarterles taxonomies non pertinentes.

Nous nous intéressons à ce stade à l'apport de la pondération des concepts des ontologies par rapport à l'affectation de concepts aux éléments feuilles ; on dispose de 51 240 éléments feuilles.

Description	**Algorithme Sans_Poids**	**Algorithme Avec_Poids**	**Amélioration due à l'algorithme Avec_*Poids***
Nombre d'éléments feuilles associés à un seul concept	12174	17555	+10.50 %
Nombre d'éléments feuilles associés à plus d'un concept	24096	13382	20.90 %

Tableau 9. Affectation des concepts aux éléments feuilles.

Au travers des résultats présentés dans le tableau 9, nous pouvons observer que :

- Un seul concept a été affecté à 12174 éléments feuilles par l'algorithme *Sans_Poids* contre 17555 par l'algorithme *Avec_Poids*. Ce qui a engendré une amélioration de 10,50 % par le second algorithme ;

- Deux ou plusieurs concepts ont été affectés à 24096 éléments feuilles par l'algorithme *Sans_Poids*, et à 13382 éléments feuilles par l'algorithme *Avec_Poids*. Comme nous cherchons à affecter un seul concept, nous pouvons alors conclure que l'algo*rithme Avec_Poids* améliore le résultat de 20.9%.

Les résultats de l'affectation des concepts aux éléments feuilles s'expliquent par le fait que, dans un élément, nous pouvons trouver à la fois un concept et son concept-fils. L'algorithme *Sans_Poids* affecte ces deux concepts à l'élément en question. Par contre, l'algorithme *Avec_Poids* retient le concept fils car la pondération automatique des taxonomies que nous avons proposée accorde plus d'importance aux concepts fils.

5.5 Conclusion

Dans ce chapitre, notre souci était de valider notre méthode construction automatique d'une structure sémantique par document XML. Pour ce faire, nous avons réalisé une expérimentation en utilisant : 1) Le thésaurus *MeSH* spécialisé dans le domaine médical, et 2) Une base de tests composée d'un échantillon de 41663 documents XML de la collection médicale *XMLImageCLEFMed 2010*.

Afin de comparer et d'évaluer l'apport de la pondération des taxonomies, nous avons réalisé : 1) Des tests en tenant compte de la pondération automatique des concepts des taxonomies (algorithme *Avec_Poids*), et 2) des tests sans tenir compte des poids des concepts (algorithme *Sans_Poids*).

Les expérimentations réalisées montrent que la pondération automatique des taxonomies a amélioré l'affectation des taxonomies aux documents ainsi que l'association des concepts aux éléments feuilles textuels dans les documents XML orientés-textes. Grace à ces résultats, on peut conclure que la structure sémantique d'un document permettra de cibler plus de documents pertinents sémantiquement et ceci soit dans un processus de recherche d'information, soit dans un processus OLAP d'analyse en ligne de documents.

Cette amélioration de résultat s'explique par le fait que notre algorithme proposé pour la pondération des concepts des taxonomies permet de sélectionner la taxonomie pertinente pour un document donné et, en conséquence, les concepts pertinents à affecter aux éléments feuilles de la structure sémantique de ce document.

Il aurait été intéressant de comparer nos résultats avec ceux des travaux de la littérature. Néanmoins, nous avons été confrontés au non disponibilité des ressources sémantiques nécessaires. De plus, cette comparaison nécessitera beaucoup d'efforts et de temps.

Conclusion générale

Le travail présenté dans ce mémoire de thèse s'inscrit dans un contexte de structuration sémantique de documents *XML* en partant de leur structure logique et de leur contenu.

Le premier chapitre était consacré à la présentation des concepts de base utilisés dans le cadre de nos travaux à savoir l'entreposage de documents XML et l'indexation des éléments textuels (indexation classique et indexation sémantique). Le deuxième chapitre présente un tour d'horizon des principaux travaux abordant la sémantique des documents en se focalisant sur l'indexation sémantique. Dans le troisième chapitre, notre intérêt s'est porté sur la définition et la proposition d'une approche pour la construction d'une structure sémantique par document *XML*. Cette approche se compose de cinq étapes :

1) Extraction des termes significatifs pour les éléments de type texte d'un document *XML* (feuilles de l'arbre XML).
2) Choix d'une taxonomie pour chaque document, celle qui correspond le mieux à sa sémantique.
3) Association, à chaque élément feuille de la structure spécifique du document, du concept significatif à partir de la taxonomie retenue.
4) Propagation des concepts aux éléments non feuilles.
5) Traitement du cas particulier des éléments de type métadonnées (sans concept associé).

Les structures sémantiques, telles que définies, seront utiles pour enrichir les analyses multidimensionnelles des documents de l'entrepôt à plusieurs titres : (a) pour effectuer des recherches et des analyses sur les éléments textuels et non plus uniquement sur des valeurs factuelles ou numériques ; (b) pour expliquer le processus d'analyse et d'aide à la décision (construction d'annotations et de descripteurs sémantiques des faits et des dimensions grâce aux concepts des ressources sémantiques associées aux documents). De fait, la structuration sémantique des documents (qu'il s'agisse de documents XML ou de tout autre type de document) est utile dans bon nombre de domaines d'application, tant en informatique décisionnelle, qu'en recherche d'information ou encore en gestion électronique des documents. Autant de domaines où la détermination de la sémantique des contenus par une indexation et une structuration sémantique peut être un apport considérable pour aider l'utilisateur à cibler au mieux les documents pertinents et en analyser les contenus dans une masse de documents grandissante.

Le quatrième chapitre s'intéresse à l'exploitation de la structure sémantique. Plus précisément, nous avons intégré les structures sémantiques dans les entrepôts de documents et nous avons introduit le concept de méta-document ; c'est ainsi que nous avons construit et testé un ensemble de requêtes analytiques de deux types (RI et OLAP) sur les documents XML en utilisant le langage de requête XQ*uery*.

Dans le dernier chapitre, nous avons mené des expérimentations pour valider notre démarche de structuration sémantique de documents *XML*. Pour cela, nous avons utilisé : 1) Le thésaurus *MeSH*, spécialisé dans le domaine médical, pour indexer et rechercher des articles, et 2) la collection médicale de documents XML *ImageCLEFMed* 2010 comme base de test.

Nous avons conclu que la pondération automatique des poids des concepts des taxonomies que nous avons proposée a amélioré les différents résultats obtenus (affectation des documents aux taxonomies ainsi que celle des éléments feuilles aux concepts).

Plusieurs perspectives à ce travail sont envisageables. Dans un premier temps, l'utilisation de la structure sémantique peut s'étendre naturellement à un processus de fouille sémantique sur les documents XML orienté-textes. Dans un deuxième temps, nous comptons étendre ces travaux pour pouvoir associer plusieurs structures sémantiques à un même document XML (multi-structuralité sémantique des documents) afin de traduire les points de vue de plusieurs lecteurs surtout lorsque le document peut appartenir à plusieurs domaines. Enfin, nous envisageons de proposer de nouveaux opérateurs pour exploiter cette structure sémantique. Comme par exemple, roll up ou drill down pour passer d'un niveau à un autre dans la structure sémantique).

Bibliographie

Bibliographie

Abascal R. (2005), "Nouveau modèle de documents pour une bibliothèque numérique de thèses accessible par leur contenu sémantique", *Thèse de doctorat*, INSA, Lyon, pp.1-236, 2005.

Aronson A. R. (2001), "Effective mapping of biomedical text to the *UMLS* Metathesaurus: the MetaMap program", *In Proceedings AMIA Symposium*, pp. 17–21.

Aronson A. R., Mork J. G., CW Gay S. M. H. et Rogers,W. J. (2004),"The NLM Indexing Initiative's Medical Text Indexer*", InMedinfo* 2004, pp. 268–272.

Aussenac G. N., Mothe J. (2004), "Ontologies as Background Knowledge to Explore Document Collections", *In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO)*, France, April 26-28, pp. 129-142.

Baziz M. (2005), "Indexation conceptuelle guidée par ontologie pour la recherche d'information", Thèse de doctorat, université de Toulouse 1, pp.1-218, France.

Baziz M., Boughanem M., Prade H. (2007), "Une approche de représentation de l'information en RI basée sur les sous-arbres", *Conférence en Recherche d'Information et Applications (CORIA 2007)*, Saint-Etienne, France, pp. 335-350.

Baeza-yates R., Ribeiro-Neto B. (1999), "Modern Information Retrieval ", 1ère edition, Addison-Wesley, ISBN 0-201-39829-X.

Baeza-yates R., Navarro G. (1994), "Integrating contents and structure in text retrieval", *SIGMOD RECORD*, 25(1), pp. 67-79.

Benoît H., Adeline N., André S. (1997), "Les ressources lexicales pour l'étiquetage sémantique", Dans le chapitre 3 de l'ouvrage "Les linguistiques de corpus", sous la direction des doctorants et maîtres de conférences de l'université de Paris XIII, pp. 1-254.

Ben Meftah S., Khrouf k., Ben Kraiem M., Feki J., Soulé-Dupuy C. (**DBLP**) (2012), "Une approche pour l'extraction automatique de structures sémantiques de documents XML", *INFormatique des ORganisations et Systèmes d'Information et de Décision (Inforsid 2012)*, pages 523-538, Montpellier, France, mai 2012. ISBN2-906855-22-7.

Ben Meftah S., Khrouf k., Feki J., Soulé-Dupuy C. (2013), "Semantic Structure for XML Document: Structuring and Pruning", International Conference on Information Technology and e-Services ICITeS' 2013, March 24-26, Sousse, Tunisia.

Ben Meftah S., Khrouf K., Feki J., Soule-Dupuy C. (**DBLP**)(2014), "Structuration sémantique des documents XML : Expérimentations et évaluation", *COnférence en Recherche d'Information et Applications (CORIA'2014)*, Nancy, France, Mars 2014, pages 53-62, 19-21, ISBN 978-2-37111-001-4.

Ben Meftah S., Khrouf K. , Feki J., Soulé-Dupuy C. (2015) , "A semantic approach for XML document warehousing and OLAP analysis", *International Journal of Information and Decision Sciences*, Inderscience edition, USA, to appear. (Accepted on 16 April 2015).

Ben Messaoud I., Feki J., Khrouf K., Zurfluh G. (2011), "Unification of XML Document Structures for DOCW", *International Conference on Enterprise Information Systems(ICEIS'11)*, pp. 85-94, Beijing, China.

Bevan K., Guido Z., Peter B., Laurianne S., Michael L. (2012), "Graph-based Concept Weighting for Medical Information Retrieval", *ADCS'12* Proceedings of the Seventeenth-Australasian Document Computing Symposium, December, Dunedin, New Zealand, pp.80-87, ISBN: 978-1-4503-1411-4.

Bodenreider O., Nelson S. J., Hole W. T. et Chang H. F. (1998), "Beyond synonymy: exploiting the *UMLS* semantics in mapping vocabularies", *Proceedings of AMIA*, pp. 815–819.

Boubekeur F., Azzoug W.(2013), "Concept-based indexing in text information retrieval".*International Journal of Computer Science & Information Technology (IJCSIT)*, Vol 5, No 1, February 2013.

Boughanem M., Kraaij W., and Nie J-Y (2004), "Modèles de langue pour la recherche d'informations, dans Les systèmes de recherche d'informations", Book Chapter : Modèles conceptuels, ed. M. Ihadjadene, Hermes-Lavoisier, pp. 163-184.

Bouillon P., Fabre C., Sébillot P., Jacqmin L. (2000), "Apprentissage de ressources lexicales pour l'extension de requêtes", T*raitement automatique des langues, numéro spécial traitement automatique des langues pour la recherche d'information Vol 41*, N°2, pp. 367–393.

Brill E. (1992), "A simple Rule-based Part-of-speech Tagger", *Proceedings of the 3rd Conference on Applied Natural Language Processing*(ANLP 1992). March 1992, Trento, Italy. pp 152-155.

Bruandet M-F., ChevalletJ-P., ParadisF. (1997), "Construction de thesaurus dans le système de recherche d'information IOTA : application à l'extraction de terminologie", *Actes des Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue (Réseau FRANCIL, JST'97)*, Avignon (France), pp. 537-544, Avril 1997.

Bruce Croft W., Jangwon S. (2008), "Blog site search using resource selection",*ACM 17th Conference on Information and Knowledge Management, CIKM 2008*, 2-6 November in Hong Kong, pp.1053-1062.

Calabretto S. (2003), "Modèles de représentation de la sémantique des documents : Applications aux bibliothèques numériques", *Mémoire d'habilitation à diriger des recherches*, Institut National des Sciences Appliquées de Lyon et l'Université Claude Bernard Lyon I, juin 2003.

Ceausu V., Despres S. (2005), "Fouille de textes pour orienter la construction d'une ressource Terminologique",Extraction et gestion des connaissances (EGC'2005), Actes des cinquièmes journées Extraction et Gestion des Connaissances, vol. RNTI-E-3, pp.239-244, Paris du 18 au 21 janvier.

Chauché J. (1984). "Un outil multidimensionnel de l'analyse du discours".*Proceedings of the 22nd conference on Association for Computational Linguistics*, July 1984, Stanford California. pp 11-15.

Cohen S., Mamou J., Kanza Y., Sagiv Y. (2003), "XSEarch: A Semantic Search Engine for XML", *Proceedings of the 29th VLDB Conference, Berlin, Germany*, pp. 45-56.

Cunningham H., Maynard D., Bontcheva K., Tablan V. (2002), "GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications", *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics,* Philadelphia, pp. 168-175, July 2002.

Cutting D., Kupiec J., Pedersen J. et Sibun P. (1992), "A practical Partof- Speech tagger", In *Proceedings of the 3rd conference on Applied Natural Language Processing*, pp. 133–140, Stroudsburg, PA, USA.

Dinh D., Tamine L. (2010), "Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients", *Conférence francophone en Recherche d'Information et Applications, CORIA 2010*, pp. 325-336*,* France Toulouse.

Dinh D. (2012), "Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques", *Thèse de doctorat*, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), pp.1-309, France Toulouse.

Ding Y., Engels R. (2001), "Using co-occurrence Theory to Generate Lightweight Ontologies", DEXA *Workshop*, pp. 961-965.

Lindberg D., Humphrey B., McCray A. (1983), "The unified medical language system", *National Library of Medicine, Bethesda, MD, in Methods of Information in Medicine*, pp. 281-291.

Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research Development, 2(2):159-165.

Egozi O., Markovitch S., and Gabrilovich E. (2011), "Concept-Based Information Retrieval using Explicit Semantic Analysis", *ACM Transactions on Information Systems*, pp.1-34, 2011.

Erdmann M., Maedche A., Schnurr H., Staab S. (2000), "From manual to semi-automatic semantic annotation : About ontology-based text annotation tools", *Proceedings* of the *COLING 2000 Workshop* on *Semantic Annotation* and *Intelligent Content*, Luxembourg, pp. 1-7, August *2000*.

Fay-Varnier C., Fouqueré C., Prigent G. et Zweingenbaum P. (1991), "Modules syntaxiques des systèmes d'analyse du français. TSI", T*echniques et Science Informatiques, Éditions AFCET-Bordas*, 1991. Volume 10, N°6, pp. 403-425.

Fellbaum C. (1998), "WordNet, an Electronic Lexical Database", with a preface by George Miller. Cambridge, MA: MIT Press; 1998. 422 p. $50.00

Fuhr N.(2000), "Models in Information Retrieval", *Lectures on Information Retrieval, ESSIR 2000, Lecture Notes in Computer Science*: M. Agosti & F. Crestani (Eds.), Springer Verlag, pp. 21-50.

Fuhr N. and Großjohann K. (2001), "XIRQL: A Query Language for Information Retrieval in XML Documents*", In Proceedings of SIGIR 2001*, Toronto.

Gale W. A., Church K. W., Yarowsky D.(1992), "One sense per discourse ", *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pp. 233-237.

Gammoudi M. (1993), "Méthode de décomposition rectangulaire d'une relation binaire : une base formelle et uniforme pour la génération automatique des thesaurus et la recherche documentaire", *Thèsede doctorat de l'Université de Nice-Sophia Antipolis,* Spécialité informatique.

Gilchrist A. (2003), "Thesauri, taxonomies and on tologies an etymological note", *Journal ofDocumentation*, Vol.59, nº. 1, 2003, pp.7–18. Disponible sur:,http://bibliologia.info/archivos/Thesauros%20taxonomias.pdf [Consulté le 28 septembre 2011].

Gruber T.R. (1993), "Toward Principles for the design of Ontologies used for Knowledge Sharing", *in Proc of International Workshop on Formal Ontology,* Padova, pp.1-23, Italy, March.

Guo L., Shao F., Botev C., Shanmugasundaram J. (2003), "XRANK: Ranked Keyword Search over XML Documents", SIGMOD, june 9-12, San Diego, CA, copyright ACM.

Haav H.M., Lubi T.L. (2001), "A Survey of Concept-based Information Retrieval Tools on the Web", *Proceedings of the 5th East-European Conference ADBIS*, Vol 2, pp. 29-41.

Hachaichi Y., Feki J., Ben-Abdallah H. (2010), "Modélisation multidimensionnelle de documents XML centrés-données", *Journal of Decision Systems*, Vol. 19(3), pp. 313-345, Edition Hermès.

Harrathi F., Calabretto S., Roussey C. (2007), "Multilingual Extraction of Semantic Indexes". *Workshop on Semantically Aware Document Processing and Indexing (SADPI 07)*, Montpellier, pp.1-11, France.

Harrathi R., Calabretto S. (2010), "Une approche de recherche sémantique dans les documents semistructurés", *Atelier Recherche d'Information Sémantique,* pp. 1-20, Marseille, France.

Hernandez N. (2005), "Ontologies de domaine pour la modélisation du contexte en Recherche d'information", *Thèse de doctorat de l'Université Paul Sabatier de Toulouse*, Spécialité Informatique, pp. 1-248.

Kang B. Y., Lee S. (2005), "Document indexing: a concept-based approach to term weight estimation", *Information Processing and Management*, Vol. 41, Issue 5, pp. 1065–1080.

Khrouf K. et Soulé-Dupuy C. (2004), "A Textual Warehouse Approach: a Web Data Repository", *Intelligent Agents for Data Mining and Information Retrieval, Idea Group Publishing*, DOI: 10.4018/978-1-59140-194-0, pp. 101-124.

Khrouf K., Soulé-Dupuy C. (2005), "DocWare : Vers l'entreposage et l'analysemultidimensionnelle de documents", *Conférence en Recherche d'Information et Application (CORIA'05)*, pp. 405-420, Grenoble, France.

Khrouf K., Feki J., Soulé-Dupuy C. (2011), "An Approach for Multidimensional Analysis of Documents", *International Conference on Information Systems and Economic Intelligence*, pp..46-53, Marrakech, Maroc.

Khrouf K., Ben Meftah S., Feki J., Ben Kraiem M., Soulé-Dupuy C. (2012), "Document warehouse: integration of semantic structure", International Conference on Information Systems and Economic Intelligence (SIIE 2012), Jerba 2012 (Among Best Papers).

Kimball R. (2002), "The Data Warehousing Toolkit second edition", Wiley, New York.

Kiryakov A., Popov B., Terziev I., Manov D. (2004), "Ognyan off D., Semantic annotation, indexing, and retrieval", *Journal of Web Semantics*, Vol 2, No 1, pp. 49-79.

Lallich-Boidin G., Maret D., "Recherche d'information et traitement de la langue : fondements linguistiques et applications",*Préf. Par S. Chambaud. Les Cahiers de l'enssib, n°3. Lyon, Les Presses de l'enssib*, pp. 1-288, ISBN: 978-2-910227-60-9.

Extensible Markup Language (XML)1.0.(2008), "Ewtensible Markup Language (XML) 1.0", Word Wide Web Consortium (W3C) Recommandation, 2008. http://www.w3.org/TR/2008/REC-xml-20081126/.

Fourel F. (1998), "Modélisation, indexation et recherche de documents structurés", *Thèse de doctorat*, Université Joseph Fourier, pp.1-251, Grenoble.

Majdoubi J., Loukil H., Tmar M. and Gregory F. (2012), "Biomedical indexing and retrieval system based on language modeling approach", *International Journal of Software Engineering & Applications (IJSEA)*, Vol.3, No.3, May 2012.

Morris J. et Hirst G. (1991), "Lexical cohesion computed by thesaural relations as an indicator of the structure of text", *Journal Computational linguistics*, Volume 17, Issue 1, March 1991, pp 21–48.

Naffakhi Najeh M. (2013), "Un modèle de recherche d'information agrégée basée sur les réseaux bayésiens dans des documents semi-structurés", *Thèse de doctorat de l'Université Paul Sabatier de Toulouse*, pp.1-143, Spécialité Informatique.

Neches S., Fikes R., Finin T., Gruber T., Patil R., Senator T. (1991), "Enabling technology for knowledge sharing", *AI Magazine*, 12, pp.36–56.

Patrick Séguéla M. (2001), "Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques", Thèse de doctorat de l'Université Paul Sabatier de Toulouse, Spécialité Informatique.

Patriarche R., Gedzelman S., Diallo G., Bernhard D.,Bassolet C.(2005), "Noesis Annotation Tool : Un outil pour l'annotation textuelle et conceptuelle de documents", *Ingenierie des Connaissances IC'2005*, pp.1-59, Nice (France) Mai 2005.

Porter M. (1980), "An algorithm for suffix stripping Program", 14(3), pp. 130-137.

Poullet L. (1997), "Formaliser la sémantique des documents : Un modèle unificateur", *INFormatique des ORganisations et Systèmes d'Information et de Décision (Inforsid 1997)*, pp. 339-352, ISBN 2-906855-13-8.

Ravat F. (2010), "Finding an Application-Appropriate Model for XML Data warehouse", Inf. Syst. 35(6), p.662-687, 2010.

Roisin C. (1999). "Documents structurés multimédia", *Habilitation à diriger des recherches*, InstitutNational Polytechnique de Grenoble.

Roussey C. (2001), "Une méthode d'indexation sémantique adaptée aux corpus multilingues", *thèse de doctorat de l'INSA de Lyon informaique*, pp.1-150, Lyon.

Salton G., McGill M.J. (1983), "Introduction to modern Information Retrieval ". *Mc Graw Hill International Book Company (New-York)*, ISBN 0-07-Y66526-5, 2ème Edition.

Sauvagnat K. (2005), "Modèle flexible pour la Recherche d'Information dans des corpus de documents semi-structurés", *Thèse de doctorat de l'Université Paul Sabatier de Toulouse*, Toulouse III, Spécialité Informatique, pp. 1-237.

Seydoux F. (2006), "Exploitation de connaissances sémantiques externes dans les représentations vectorielles en recherche documentaire", *thèse en informatique*, Ecole polytechnique fédérale de LAUSANNE, pp.1-160.

Sophia A., Richard G. (2006), "APIs au-dessus de JDBC", Université de Nice, Version 0.3, pp 1-4.

Soulé-Dupuy C. (2001), "Bases d'informations textuelles: Des modèles aux applications", *Mémoire d'HDR*, Université Paul Sabatier, Toulouse III, Décembre 2001.

Sparck-Jones K., Robertson S.E., Hiemstra D., Zaragoza H. (2003), "Language modeling and relevance", *Language Modeling for Information retrieval, Kluwer Academic Publishers*, pp. 57-71.

Sullivan D. (2001), "Document Warehousing and Text Mining", *Wiley Edition*, Université du Michigan, ISBN 0471399590, pp. 1-560.

Tagarelli A., Greco S. (2010), "Semantic clustering of XML documents", *ACM Transactions onInformation Systems (TOIS)*, Volume 28, Issue 1, January 2010.

Tournier R. (2007), "Analyse en ligne (OLAP) de documents". *Thèse de doctorat, Université Paul Sabatier (Toulouse III), pp.1-204.*

Trotman A, and Sigurbjörnsson B. (2005), "NEXI, Now and Next", *In: Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval*, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004 .

Vallet D., Fernández M., Castells P. (2005), "An Ontology-Based Information Retrieval Model", *Proceedings of the 2nd European Semantic Web Conference*, pp. 455-470.

W3C (1999), "XML Path Language (XPath) Version 3.0", W3C Recommendation, 16 November 1999, disponible sur : http://www.w3.org/TR/xpath-30/. Dernier consultation: 18-04-2015.

W3C (1999), "Extensible Markup Language (XML) 1.0 (Fifth Edition)", W3C Recommendation 26 November 2008, disponible sur : http://www.w3.org/TR/REC-xml/. Dernière consultation : 26-12-2015.

Wilbur J. (2003), "PubMed Related Citations Algorithm", *Rapport technique*, Proceedings of the 2004 AMIA Symposium, pp 319-23.

XQuery 1.0 (2007),"XQuery 1.0: An XML Query Language:W3C Recommendation",disponible sur : http://www.w3.org/TR/xquery/.

XPath 2.0.(2007), "XML Path Language (XPath) 2.0. XML Path Language (XPath) 2.0: W3C Recommendation", disponible sur:http://www.w3.org/TR/xpath20/.

Xu Y., Papakonstantinou Y. (2005), "Efffcient Keyword Search for Smallest LCAs", in XMLDatabases. SIGMOD Conference, pp. 537-538.

Zargayouna H., Salotti S. (2004), "SemIndex: a model of semantic indexing on XML documents", *European Conference on Information Retrieval (ECIR'2004)*, Sunderland, UK, Vol 2, Avril 2004.

Zhou X., Zhang X., Hu X. (2006), "MaxMatcher : Biological Concept Extraction Using Approximate Dictionary Lookup", *In PRICAI*, volume 4099, pp.1145–1149.

Annexe : Exemple d'un document XML

A Novel Multidimensional Model for the OLAP on documents: Modeling, Generation and Implementation

Maha Azabou[1], Kaïs Khrouf[1], Jamel Feki[1], Chantal Soulé-Dupuy[2], Nathalie Vallès[2]

University of Sfax, Faculty of Economics and Management, Computer Department, MIR@CL Laboratory, University of Sfax
Airport Road Km 4, P.O. Box. 1088, 3018 Sfax, Tunisia
Azabou.Maha@yahoo.fr, Khrouf.Kais@isecs.rnu.tn, Jamel.Feki@fsegs.rnu.tn
[2] IRIT, University of Toulouse 1 Capitole,
2, Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9, France

{Chantal.Soule-Dupuy, Nathalie.Valles-Parlangeau}@ut-capitole.fr

Keywords:XML Documents, OLAP, *Diamond* multidimensional model.

Abstract :

As the amount of textual information grows explosively in various kinds of business systems, it becomes more and more essential to analyze both structured data and unstructured textual data simultaneously. However infor-mation contained in non structured data (documents and so on) is only partially used in business intelligence (BI). Indeed On-Line Analytical Processing (OLAP) cubes which are the main support of BI analysis in decision support systems have focused on structured data. This is the reason why OLAP is being extended to unstructured textual data. In this paper we introduce the innovative "Diamond" multidimensional model that will serve as a basis for semantic OLAP on XML documents and then we describe the meta modeling, generation and implementation of a the *Diamond* multidimensional model.

[2]Conférence: 4th International Conference on Model & Data Engineering (MEDI'2014)

Liste des publications de la thèse

Articles de journaux

7. Ben Meftah S., Khrouf k., Feki J., Soulé-Dupuy C., "Semantic Structure for XML Documents: Structuring and Pruning", Journal of Information Organization (http://www.dline.info/jio/v3n1.php), page 37-46, volume 3, issue 1, Print ISSN 2278 6503, Online ISSN2278 – 6511, 2013.

8. Ben Meftah S., Khrouf K. , Feki J., Soulé-Dupuy C. , "A semantic approach for XML document warehousing and OLAP analysis", International Journal of Information and Decision Sciences, Inderscience edition, USA, to appear. (Accepted on 16 April 2015).

Articles de conférences

1. Ben Mefteh S., Feki J., Hachaichi Y., "Élaboration de schémas de magasins de données à partir d'une base de données objet", *Atelier des Systèmes Décisionnels (ASD'08),* pages 29-40 Mohammedia-Maroc, 10-11 Novembre 2008.

2. Ben Mefteh S., Feki J., Hachaichi Y., "CAME-BDO: A tool for designing data marts from object databases", *International Arab Conference on Information Technology (ACIT'10),* Garyounis, Libya, December 2010.

3. Khrouf k., Feki J., Ben Kraiem M., Soulé-Dupuy C., "Document warehouse: integration of semantic structure", *International Conference on Information Systems and Economic Intelligence (SIEE 2012)*, Jerba 2012 (**Among Best Paper**).

4. Ben Meftah S., Khrouf k., Ben Kraiem M., Feki J., Soulé-Dupuy C. (**DBLP**), "Une approche pour l'extraction automatique de structures sémantiques de documents XML", *INFormatique des ORganisations et Systèmes d'Information et de Décision (Inforsid 2012)*, pages 523-538, Montpellier, France, mai 2012. ISBN2-906855-22-7.

5. Ben Meftah S., Khrouf k., Feki J., Soulé-Dupuy C., "Semantic Structure for XML Document: Structuring and Pruning", *International Conference on Information Technology and e-Services ICITeS' 2013,* March 24-26, Sousse, Tunisia.

6. Ben Meftah S., Khrouf K., Feki J., Soule-Dupuy C. (**DBLP**), "Structuration sémantique des documents XML : Expérimentations et évaluation", *COnférence en Recherche d'Information et Applications (CORIA'2014)*, Nancy, France, Mars 2014, pages 53-62, 19-21, ISBN 978-2-37111-001-4**.**

RÉSUMÉ

L'indexation sémantique représente les documents par des concepts qui reflètent le contenu des documents au lieu que par des termes simples (aussi appelés « sacs de mots »). L'objectif de cette thèse est de proposer une approche pour la construction d'une structure sémantique des documents XML centrés-textes et ceci à partir de leur structure logique et de leur contenu. La construction de cette structure sémantique passe par trois étapes principales, à savoir : 1) La détermination d'une taxonomie à choisir parmi un ensemble de taxonomies de domaine et son affectation au document. Ce choix de la taxonomie se base sur une démarche de pondération des concepts des taxonomies existantes ; dans cette pondération nous favorisons les concepts les plus spécifiques assumant qu'ils sont plus porteurs de sens que leur ascendant ; 2) L'affectation d'un concept significatif de la taxonomie retenue à chaque élément feuille de la structure spécifique du document ; et 3) L'inférence des concepts aux éléments non feuilles de la structure à travers un ensemble de règles. Dans le but de montrer l'utilité de la structure sémantique, nous avons définit un ensemble de requêtes Query sous foorme classique et OLAP. Afin d'évaluer notre approche de construction de structure sémantique, nous avons réalisé des expérimentations en utilisant : 1) le thésaurus MeSH spécialisée dans le domaine médical et 2) la collection de documents XMLImageCLEFmed 2010. Ces expérimentations sont réalisées dans un contexte de recherche d'information (RI) et aussi dans des traitements analytiques en ligne (OLAP).

Mots-clefs

Document XML, Structure Sémantique, Taxonomie, Indexation sémantique, RI, OLAP.

Summary

The semantic indexation represents documents by concepts that better reflect the content of the documents than single terms. The objective of this work is to propose an approach for building a semantic structure for text-centric XML documents; this semantic structure construction relies on the logical structure and content of the document. This semantic structuring includes three main steps, namely: 1) The determination of a taxonomy, among a set of domain taxonomies, that will be assigned to the document; this determination is based on a weighted-approach of taxonomies, which favor the specific concepts assuming them more meaningful than their ascendant; 2) The assignment of a significant concept of the retained taxonomy, to each leaf-element of the specific structure of the document; and 3) The inference of concepts to non-leaf nodes of the semantic structure by using a set of rules. In order to evaluate our proposed approach for constructing semantic structures, we performed experiments while using: 1) The *MeSH* thesaurus, specialized in the medical domain and 2) the *XMLImageCLEFMed 2010* collection of XML documents. These experiments are realized in the context of IR and in On-Line Analytical Processing.

Keywords

XML Document, Semantic Structure, Taxonomy, Semantic indexation, IR, OLAP.