

dF.S.D ... N° D'ordre :

Université Saâd DAHLAB de Blida**Faculté des sciences****Département d'Informatique**

Mémoire présenté par :

M^{lles} Marrouqui Imane et Zemieche Manel

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique**Filière : Informatique****Spécialité : Ingénierie des logiciels**

Sujet :

**Cartographie sémantique et classification des notices
pharmaceutiques**

Soutenu le : -- Juillet 2013, devant le jury composé de :

Dr. **BENBLIDIA Nadjia**

Promotrice

Mlle. **MEZZI Melyara**

CO-promotrice

Mr. Hamouda

Président de jury

Mme. Farah

Examineur

Mme. Boumahdi

Examineur

Dédicaces

Le présent travail est pour moi une autre occasion et un agréable devoir d'exprimer mes reconnaissances et ma gratitude envers Dieu « l'omnipotent ».

Toute ma reconnaissance à mes parents pour leurs soutiens et abnégation tout au long de mon cursus en quête du savoir, j'espère que les résultats que j'obtiendrai seront enfin à la hauteur de leurs espérances.

A mes soeurs Meriem et Chanez, à mon beau-frère Mustapha, surtout à mes adorables nièces Naila et Sérine je dis, persévérez dans vos vies, ce n'est pas toujours facile, c'est vrai mais la joie de la réussite n'en sera que plus grande car étant méritée.

Mes très cordiaux remerciements à tata Leila, une femme d'exception mais aussi un modèle de la réussite dans l'accomplissement du savoir... Mon enrichissement dans le domaine « pharmaceutique » le lui doit beaucoup.

Je salue respectueusement mes professeurs pour leur patience et leur dévouement dans l'accomplissement de leur travail.

A la mémoire de mes grands-parents, papi et grand-mère Allah Yarhamhoum.

A Mami qui a toujours été ma source de volonté dans la vie.

Je n'oublie pas de rendre un grand hommage à tous mes oncles, tantes, tontons et tatas sans oublier mes cousins et cousines.

A ma meilleure amie Ouardia (et son mari Mohamed) qui s'est toujours exceptionnellement démarquée avec son amitié envers moi, sans oublier ses parents (tonton Saïd et tata Djazira) que j'ai toujours considéré comme étant ma deuxième famille.

A tous mes amis (Alister, Lamine, Djihad, David, Shinzo, Peffect et Louisa) pour leur soutien, présence et compréhension. A mes collègues, Fethi, Dahmane et Fodhil.

Une particulière dédicace à Zemieche Manel mon binôme qui a toujours su me supporter durant tout le temps de la réalisation de notre projet ensemble. A tous ceux-là je dédie ce mémoire de fin d'études.

Imènus

Dédicaces

Je dédie ce travail :

*À l'être le plus cher de ma vie Qui m'a soutenue durant tout mon parcours,
qui m'a aidée et encouragée sans cesser de garder toujours espoir
À ma chère MAMAN Rachida.*

*À l'homme qui m'a toujours guidé vers le droit chemin avec son amour, ses
sacrifices, encouragements et son soutien moral
À mon cher PÈRE Abd Esselam*

*À mes chers frères
Walid et Imad*

*À ma chère sœur
Lilya*

*Je le dédie particulièrement
À Gacem Nassim et à toute sa famille.*

À mes grands-mères Zilokha et Hbila

À toute la famille Zemieche et la famille Belhadri

À mes oncles et tantes et cousins et cousines

*À mon binôme Imane et sa famille pour les quels je souhaite une vie pleine de
joie et de réussite.*

*À mes très chers amis (es) :
Mezzora, Sarah, Amel, Abdou, Aghilas*

*À mes très chères cousines :
Ahlem Amina Amina Assia chaima Ilham Loubna Meriem Nahla Rima
Sabrina Samah Siham Souad Yasmin*

À mes très chères Dina Nora Lina (Nounoucha)

Enfin, à tous ceux que je porte dans mon cœur.

Zemieche Manel

Remerciement

En préambule à ce mémoire nous remerciant ALLAH qui nous a aidé et donné la patience et le courage durant ces longues années d'étude.

La personne que nous tenons à remercier est notre Co-promotrice M^{elle} Mezzi Malyara, pour l'orientation, la confiance, la patience qui a constitué un apport considérable sans lequel ce travail n'aurait pas pu être mené à bon port. Qu'elle trouve dans ce travail un hommage vivant à sa haute personnalité.

Nous souhaitons adresser nos remerciements les plus sincères à notre promotrice Dr. Benblidia pour avoir cru en nous en acceptant d'être notre promotrice, et de collaborer avec nous dans l'aboutissement de ce modeste travail en nous éclairant le chemin de la recherche.

Nous tenons à exprimer nos sincères remerciements à Mr. Cherif zahar et Mr Ait akache qui nous ont enseignés et nous ont soutenu dans la poursuite de nos études. Ainsi qu'à l'ensemble du corps professoral du département d'Informatique.

Nos remerciements s'étendent également aux personnes qui nous ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire.

On n'oublie pas nos parents pour leur contribution, leur soutien et leur patience.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire.

Merci à toutes et à tous.

Sommaire

Introduction générale

1. Contexte général de travail.....	6
2. Problématique.....	7
3. Objectifs de travail.....	8
4. Organisation du mémoire.....	8

Chapitre I: Généralités sur les documents textuels

1. Introduction.....	10
2. Le document électronique.....	10
2.1. Qu'est-ce qu'un document ?.....	10
2.1.1. La communication par le texte.....	10
2.1.2. Les documents électroniques.....	11
3. Opérations sur les documents.....	11
3.1. Indexation des documents électroniques.....	11
3.2. La classification automatique.....	12
3.3. La catégorisation du document.....	13
3.4. Résumé automatique.....	13
3.5. Cartographie des connaissances.....	14
4. Technologies linguistiques.....	14
4.1. La dimension sémantique dans le TAL.....	16
5. Conclusion.....	16

Chapitre II: La cartographie, généralités et méthodes

1. Introduction.....	17
2. La connaissance.....	17
3. Gestion des connaissances.....	18
4. Ingénierie des connaissances.....	19
5. La cartographie sémantique.....	19
6. Procédé d'une cartographie d'information.....	20
7. Objectifs de la cartographie d'informations.....	20
8. Approches de cartographie de connaissances.....	21
8.1. Approche par le processus.....	21
8.2. Approche par les domaines.....	22
9. Types de cartographie et leurs outils.....	24

10.	Synthèse et critiques.....	27
11.	Conclusion.....	28

Chapitre III: Spécificités des notices pharmaceutiques

1.	Introduction.....	29
2.	Les notices pharmaceutiques.....	29
2.1.	Qu'est-ce qu'une notice ?.....	29
2.2.	Format d'une notice.....	30
3.	Généralité sur la pharmacologie.....	30
3.1.	Qu'est-ce qu'un médicament.....	30
3.2.	Médicaments essentiels.....	31
3.3.	Spécialités et génériques.....	31
3.4.	Les familles de médicaments.....	32
4.	Conclusion.....	34

Chapitre IV: Analyse des besoins et modélisation du système

1.	Introduction.....	36
2.	Constat.....	36
3.	Processus global d'une cartographie sémantique.....	37
3.1.	Analyse terminologique.....	38
4.	Architecture du système.....	43
5.	Conclusion.....	44

Chapitre V: Implémentation du système

1.	Introduction.....	45
2.	Environnement de développement.....	45
2.1.	Langage utilisé.....	45
2.2.	OUTILS.....	45
3.	Diagramme d'accessibilité du système CSCNPharm.....	46
4.	Quelques algorithmes.....	47
4.1.	Cartographie.....	47
4.2.	Extraction.....	48
4.3.	Normalisation.....	48
4.4.	Recherche famille.....	48
4.5.	Construction Ontologie.....	49
5.	Quelques captures d'écran.....	49

5. Tests et validation du système	53
5.1. Corpus de tests	53
5.2. Mesures de performance utilisées.....	54
1. Définition des mesures de performance utilisées.....	54
5.3. Jeux de données.....	55
6. Conclusion	57
1. Conclusion	58
2. Perspectives.....	59
Bibliographie.....	8

Liste des acronymes

<i>SGC</i>	<i>Système de Gestion des Connaissances</i>
<i>SBC</i>	<i>Système à Base de Connaissances</i>
<i>SE</i>	<i>Système Expert</i>
<i>DCI</i>	<i>Dénomination Commune Internationale</i>
<i>TAL</i>	<i>Traitement Automatique de la Langue</i>

Les figures

Figure 1: Arbre de connaissances	23
Figure 2: Processus global de cartographie	37
Figure 3: Analyse terminologique	40
Figure 4: Architecture de transformation de vues.....	42
Figure 5: quelques formes arborescentes.....	42
Figure 6: Architecture du système CSCNPharm	44
Figure 7: Diagramme d'accessibilité du système CSCNPharm.....	47
Figure 8: Fenêtre principale de CSCNPharm	49
Figure 9: Fenêtre d'ouverture d'une notice	50
Figure 10: Affichage de la notice	50
Figure 11: Etape de segmentation.	51
Figure 12: Etape d'extraction.....	51
Figure 13: Etape de normalisation.....	51
Figure 14: Fenêtre des statistiques.....	52
Figure 15: Visualisation graphique (taxonomie).	52
Figure 16: Classification d'une notice	53

Les tableaux

Tableau 1: Types et outils de cartographie.....	27
Tableau 2: Familles des médicaments	34
Tableau 3: Classification manuelle / CSCNPharm (test et validation).	56
Tableau 4: Mesures de performances de CSCNPharm.....	56

RESUME

La quantité d'informations et de documents disponibles de nos jours, entraîne une «surinformation» de l'utilisateur final (entreprise, organisme, individu, etc.). Les utilisateurs –submergé par une masse considérable de documents- ne sont donc plus capables d'analyser ou d'appréhender les informations dans leur globalité. L'information utile étant enfouie dans le texte, il devient indispensable de proposer de nouveaux systèmes permettant de mettre en œuvre des méthodes permettant d'analyser les contenus des documents, de les organiser et de les représenter automatiquement de manière à faciliter leur compréhension et gestion.

La cartographie (i.e., visualisation) de l'information propose une solution originale et efficace au problème de prise de connaissance d'un ensemble d'information, la conversion de l'information en une représentation visuelle tirant au mieux partie des capacités de perception des individus. Dans cette optique, nous proposons une méthode originale pour la cartographie sémantique des documents textuels (plus particulièrement, les notices pharmaceutiques), et ce, en se basant sur les avantages qu'offre les ontologies au niveau formalisation des connaissances.

Mots clés : Documents Textuels, Cartographie sémantique, Classification, Ontologies, et Ingénierie des connaissances.

ABSTRACT

The amount of information and materials available today, leads to an "overload" of the end user (business, organization, individual, etc...). Users -Overwhelmed by a huge mass of material- are no longer able to analyze or understand the information in its entirety. Since the useful information is buried in the text, it is essential to introduce new systems to implement methods that analyze the contents of the documents, organize and represent them automatically to facilitate their understanding and management.

Information cartography (ie, visualization) offers a unique and effective solution to the problem of getting to know a set of information, the conversion of information into a visual representation taking the best part of the collection capacity individuals. In this context, we propose a novel method for the semantic cartography of textual documents (particularly pharmaceutical leaflets), based on the benefits of Ontologies in knowledge formalization.

Keywords: Text documents, Semantic cartography, Classification, Ontologies, and Knowledge engineering.

ملخص

كمية المعلومات والوثائق المتوقعة في يومنا هذا أدت إلى "الحمل الزائد للمعلومات" لدى مستخدم النهائي (شركة عملة، منظمة، أو فرد الخ). تطغى على المستخدمين مجموعة كبيرة من الوثائق التي تضعف قدرتهم على تحليل أو فهم المعلومات في مجملها. بما أن المعلومات المفيدة تكون مدفونة في النص، فإنه من الضروري إدخال أنظمة جديدة لتنفيذ طرق لتحليل محتويات الوثائق، تنظيمها وتمثيلها التلقائي لتسهيل فهمها وإدارتها.

رسم خرائط المعلومات (أي التصور) يقدم حلاً فريداً وفعالاً لمشكلة صنع المعرفة حول مجموعة من المعلومات، تحويل المعلومات إلى تمثيل يسمح بالاستفادة من المهارات الإدراكية للأفراد. في هذا السياق، نقترح طريقة جديدة لرسم الخرائط الدلالية للوثائق النصية (و بشكل خاص منشورات الأدوية). وذلك بالإعتماد على فوائد الانطولوجيا في إضفاء الطابع الرسمي على المعرفة.

الكلمات الرئيسية: الوثائق النصية، الرسم الدلالي الخرائط، التصنيف، الانطولوجيا وهندسة المعرفة.

Introduction générale

1. Contexte général de travail

L'homme est doté d'une capacité à visualiser l'information très développée qui joue un rôle majeur dans ses processus cognitifs¹ (reconnaissance rapide de motifs, couleurs, formes et textures). Il utilise des méthodes graphiques afin de mieux appréhender des notions abstraites ou pour représenter le monde qui l'entoure.

La cartographie de l'information relève à la fois de la visualisation scientifique², du datamining³, de l'interface homme machine, et de l'imagerie. Il s'agit de représenter dans un espace physique sous forme graphique une information souvent abstraite. Cette information peut comprendre des données, des processus, des relations ou des concepts. Sa représentation nécessite de manipuler des entités graphiques (points, lignes, formes, images, texte, surface) et leurs attributs (couleur, intensité, taille, position, forme, mouvement). Par ailleurs, on peut dire que la cartographie des connaissances est un Système de Gestion des Connaissances qui peut être vu comme une identification du patrimoine de connaissances qui permet aux organisations désireuses de gérer leur patrimoine de connaissances, d'en faire une analyse fine afin de déterminer, dans leur stratégie, quelles sont les connaissances qu'elles doivent pérenniser, développer, abandonner, etc. La cartographie devient alors un outil d'aide à la décision.

Ces dernières années, les avancées technologiques en matière d'analyse et de visualisation des informations ont conduit à la multiplication de ce type de SGC dans les organisations. Ces SGC apparaissent désormais aux yeux de nombreux utilisateurs comme des outils simples à concevoir et à utiliser, permettant une allocation des connaissances objective et rationnelle. Pourtant, ces cartographies sont elles-mêmes le produit d'une certaine vision de l'entreprise et l'expression d'un schéma cognitif individuel (d'un dirigeant) ou collectif (d'un groupe d'acteurs). Elles sont fortement empreintes d'ambiguïté ce qui conduit à s'interroger sur l'efficacité de la prise de décision qui découle de leur utilisation.

L'apparente simplicité de conception et d'utilisation d'une cartographie des connaissances comme l'apparente objectivité des analyses qu'elles permettent d'élaborer conduisent-elles à des décisions pertinentes en matière d'allocation de ressources ou conduisent-elle à renforcer des illusions sur ce que sont les

¹ Processus de compréhension et de mémorisation

² Utilisation d'images afin de comprendre les données d'origine de mesures ou de simulation

³ Gestion et exploitation des données

connaissances de l'organisation et sur les modes de gestion qu'il est possible de leur attribuer ?

2. Problématique

La cartographie sémantique des connaissances apparaît donc comme un processus délicat, qui suppose au minimum une bonne compréhension du contenu textuel où la sémantique est souvent abstraite ou cachée. Pour le moment, la diversité prévaut dans ce domaine, comme le démontre la variété des outils de construction de cartographie et de visualisation sémantique. Mais concrètement, et si on regarde le problème de plus près, on se rend compte que la construction de telles visualisations à partir d'un contenu purement textuel soulève quelques problématiques, parmi les quelles, on cite :

- Emplacement des informations : les informations peuvent être locales ou distantes. Ce qui pose des problèmes de disponibilité, identification, répartition sur plusieurs sources, variabilité des formats (encodage des caractères et description du contenu),
- Nature des documents : les documents textuels électroniques peuvent être structurés comme les bases de données, semi structurés comme les fichiers annotés, non-structurés comme les fichiers textuels dont le traitement constitue l'un des principaux problèmes du Traitement Automatique des Langues.
- Compréhension des documents textuels : Pour comprendre ce que signifie un texte ou une partie d'un texte (mot, proposition, phrase, etc.), il faut mettre en œuvre plusieurs niveaux d'analyse :
 - Lexical: signification des mots pris isolément,
 - Syntaxique: signification des associations de ces mots,
 - Contextuel: connaissance apportée par le texte à l'intérieur duquel se situe la partie de texte à analyser,
 - Pragmatique: ensemble des connaissances sur le monde auquel le texte se réfère.
- Problématiques liées à la langue : A la différence des langages artificiels, la langue naturelle est floue et équivoque elle est caractérisée par son implicité, sa redondance, et son ambiguïté.

3. Objectifs de travail

L'objectif de ce travail est de concevoir et de réaliser un système de cartographie et de classification de documents textuels de type notice pharmaceutique ayant recours aux Ontologies pour apporter une certaine uniformité et formalisation du corpus de notices, le travail vise à présenter les connaissances métiers contenu dans les notices pharmaceutiques de manière à ce qu'un non expert du domaine (un stagiaire, un vendeur, etc.) puisse les analyser et tirer des conclusions dans un but décisionnel. La cartographie des connaissances permet comme nous l'avons dit de représenter un contenu textuel sous une forme graphique facilement compréhensible par l'être humain. Par ailleurs, l'alternative Ontologique que nous avons utilisé par rapport aux autres techniques de cartographie nous assure que les nouvelles représentations soient aussi compréhensible et facilement manipulable par la machine. La représentation finale d'une notice créée par le système final doit satisfaire les caractéristiques suivantes :

- non ambiguïté,
- établie sur des distinctions fondamentalement appropriées,
- consistante et cohérente dans l'ensemble,
- possédant un niveau suffisant et adéquat de détail (pas de surcharge ou d'exagération) nécessaire à l'activité décisionnelle,
- Enfin, la granularité de l'ontologie finale se devra d'être homogène.

4. Organisation du mémoire

Afin d'atteindre les objectifs cité ci-dessus, notre mémoire s'articulera autour de cinq chapitres:

- Chapitre I: Généralités sur les documents textuels, qui constitue une porte d'entrée et une mise en contexte global sur le monde des documents textuels numériques. Nous y aborderons la définition, les différentes opérations qu'on peut appliquer sur les documents, ainsi que les technologies linguistiques de traitement automatique de la langue.

- Chapitre II : La cartographie, généralités et méthodes, dans lequel nous présenterons une définition de gestion et de l'ingénierie des connaissances, le principe de la cartographie sémantique, ainsi que les différentes techniques, méthodes, et outils de visualisation. Nous concluons ce chapitre par une synthèse critique de ces dernières.

– Chapitre III : Spécificités des notices pharmaceutiques, qui portera sur les spécificités liées aux notices pharmaceutique et les différents types de médicaments qu'on peut trouver.

– Chapitre IV: Analyse des besoins et conception, dans lequel, nous parlerons de l'approche proposée suite à l'étude bibliographique, nous présenterons la conception du système suivant. Partant de l'analyse des besoins fonctionnels et techniques, ensuite, l'analyse du système et présentation des scénarios et enfin, la présentation de l'architecture statique.

– Chapitre V: Implémentation du système, dans lequel nous présenterons l'environnement de développement (langages et outils utilisés). Ensuite, le diagramme d'accessibilité, l'architecture de déploiement du système et enfin les impressions d'écran.

En fin, la conclusion de ce mémoire synthétisera nos principales contributions et donnera quelques perspectives à notre travail.

Chapitre I :
Généralités sur les
documents
textuels

1. Introduction

Le passage du XX^{ème} au XXI^{ème} siècle a été marqué par deux phénomènes importants [1] : Le premier est d'ordre technologique « *la généralisation du document numérique* » et le second est d'ordre économique « *la reconnaissance du fait que la compétitivité économique réside dans la maîtrise des flux d'information* ». Ces deux phénomènes ont rendu indispensable le développement d'outils de traitement de l'information et en particulier de ce qui représente plus de 80 % de celle-ci, l'information textuelle. Dans ce chapitre, nous allons présenter quelques définitions rudimentaires relatives aux documents textuels.

2. Le document électronique

2.1. Qu'est-ce qu'un document ?

Le terme document [2], provenant du latin «documentum», signifie «pièce écrite servant d'information, de preuve» ou «objet quelconque servant de preuve, de témoignage». Le document en tant que medium a vu sa définition modifiée au fil du temps. Le document était à l'origine un enregistrement d'un discours oral par le biais d'un codage, le texte. Aujourd'hui, il peut être simplement défini comme un support physique ou numérique d'information.

2.1.1. La communication par le texte

La communication par un texte [2] nécessite un émetteur, celui qui envoie un message, et un destinataire, celui qui le reçoit. Afin que destinataire et destinataire se comprennent, on doit émettre une hypothèse de connaissance mutuelle (mutual-knowledge hypothesis) (Gibbs, 1987), c'est à dire que l'interprétation du message faite par le destinataire et celle voulu par le destinataire peut correspondre.

Un texte est composé d'unités linguistiques ordonnées pour que ce dernier ait un sens. D'un texte émergent des attributs et des caractéristiques à deux niveaux: le niveau lexico-syntaxique (micro) et le niveau structurel (macro). L'analyse «micro» du texte consiste à étudier ce texte d'un point de vue lexical et syntaxique. L'analyse «macro» quant à elle consiste à observer le texte comme une structure de segments.

2.1.2. Les documents électroniques

Le document est un moyen de communication interpersonnelle et sociale entre son rédacteur et le lecteur [2]. Le rédacteur utilise le document pour décrire et synthétiser ses objectifs de communication. Pour cela, il doit faire en sorte que le document réponde aux attentes du lecteur.

Le document électronique tend à se substituer au document papier. Un document électronique est un objet informatique manipulable sur un ordinateur ou sur un appareil électronique. Les documents électroniques [2]:

- sont facilement manipulables par le créateur qui prend à son compte les avantages de l'outil informatique (copier/coller, contenu dynamique, etc.) ;
- peuvent comporter des liens internes ou externes pour incorporer d'autres medias (textes, images, vidéos, etc.) ;
- sont indépendants du support;
- sont transportables et transmissibles (par réseau) ;
- sont duplicables ;
- La recherche d'objet (par exemple, un mot dans un texte) est possible au sein d'un document électronique.

C'est sur cette dernière caractéristique que reposent les SRI (Système de Recherche D'Information). La recherche d'un objet dans un document va permettre de retenir ce document si l'intention de l'utilisateur est d'obtenir un document contenant cet objet.

3. Opérations sur les documents

3.1. Indexation des documents électroniques

L'accès aux documents numériques volumineux ou complexes peut être facilité par un index du style que l'on retrouve à la fin d'un livre [3], présentant schématiquement les concepts abordés dans le document et les liens que l'auteur a établi entre eux. Il peut s'avérer un outil précieux dans la fouille de documents.

Du point de vue de la documentation, l'indexation est une opération intellectuelle impliquant une analyse approfondie d'un document et la représentation condensée de l'information portée par ce document. Le processus d'indexation permet de résoudre les problèmes et ambiguïtés du langage naturel.

L'informatique considère l'indexation comme un repérage des informations dans un ensemble de documents [1], opération qui permet d'accélérer le processus de recherche de l'information. La première génération des systèmes dits d'indexation automatique sur le texte intégral était fondée sur la création d'index (fichiers inversés et fichiers topologiques) permettant de cibler un terme ou un ensemble de termes au sein d'un corpus déterminé.

L'indexation des documents [4] était pour longtemps réalisée par des documentalistes experts du domaine, ce qui garantissait un traitement de bonne qualité, mais ça n'était pas sans poser certains problèmes relatifs au coût et à la cohérence à long terme. Le traitement de volumes de données importants est aussi un point critique car le passage à une grande échelle (scalability) représente un obstacle.

3.2. La classification automatique

Outre la construction d'un index, le processus d'extraction des termes ou des syntagmes peut permettre la construction automatique d'agrégats de termes. La classification est un processus qui consiste à construire automatiquement des classes de mots (appelées aussi agrégats, ou clusters en anglais) à partir des mots qui sont conceptuellement proches dans le texte.

Cette approche ascendante (bottom up) de l'organisation et de la représentation des connaissances correspond à une tradition épistémologique fondée sur une approche inductive. Elle réfute en effet l'hypothèse qu'il est possible d'organiser rationnellement les objets de connaissance a priori. Dans le domaine de l'analyse textuelle, elle organise les documents en fonction des occurrences lexicales qu'ils contiennent et non en fonction d'un plan de classement préexistant.

La construction automatique d'agrégats donne lieu à des représentations diverses. Une représentation fréquente des clusters se fait en particulier sous forme d'une visualisation de l'information, parfois sous forme de cartes ou de graphes, dont l'objectif est d'aider à appréhender et analyser rapidement un important volume d'informations textuelles.

3.3. La catégorisation du document

À l'inverse de la classification, la catégorisation consiste à classer des textes en fonction d'un ensemble préexistant de catégories structurées, organisées et éventuellement hiérarchisées. Ce traitement, principalement fondé sur une identification des termes du document, vise à assigner automatiquement un document ou un flux entrant d'informations textuelles dans le plan de classement préexistant, souvent construit manuellement. Cette approche correspond à une tradition épistémologique ancienne (arbre de Porphyre, tradition encyclopédique, etc.) qui présuppose l'existence d'un modèle conceptuel d'organisation rationnelle du monde formalisé dans des classes. Ce modèle peut s'exprimer à travers différents outils comme les répertoires, les thésaurus, les réseaux sémantiques, les *ontologies* et plus récemment les approches connues sous le nom de web sémantique et topic maps. La caractéristique commune de ces outils est de procéder selon une approche déductive ou top down, c'est-à-dire que les classes sont « projetées » sur les documents.

De nombreux logiciels de recherche et d'analyse de l'information, notamment en *contexte* de veille, proposent cette fonctionnalité qui constitue une aide efficace au classement automatique des flux informationnels entrants.

3.4. Résumé automatique

Alors que l'indexation vise à décrire le contenu d'un texte au moyen de descripteurs, résumer un texte consiste [5] à produire une description textuelle de son contenu en appliquant un taux variable de réduction. Un rapide examen des résumés proposés par les systèmes commerciaux montre que la frontière entre ces deux modes de représentation est parfois très floue.

Ainsi, on distingue divers types de résumés en fonction de l'usage qui en sera fait. On a notamment le résumé informatif qui donne une information générale sur le contenu d'un texte en reprenant les éléments essentiels de celui-ci, le résumé indicatif qui couvre l'ensemble des thèmes développés dans le texte et qui sert de « point d'entrée » au texte sans se substituer à lui, le résumé critique, le résumé de conclusions, etc. La notion de résumé est donc ambiguë.

À l'heure actuelle, les « résumés » produits automatiquement sont essentiellement des extractions d'unités linguistiques jugées représentatives. On ne

peut guère les considérer comme de véritables systèmes de résumé automatique au sens linguistique du terme, mais ces outils donnent des résultats intéressants en fournissant des « clés de lecture » pour l'accès au texte. Face au volume croissant d'informations, ces outils permettent de prendre rapidement connaissance d'un texte volumineux ou d'un ensemble de textes dans le cas des systèmes de résumé multi-documents.

3.5. Cartographie des connaissances

Les cartographies des connaissances [6] sont apparues relativement récemment dans les entreprises, concomitamment au développement des Systèmes de Gestion des Connaissances mais également au développement des technologies d'analyse et de visualisation des informations. Ces Knowledge Maps, Knowledge Mapping, Knowledge Cartography, Knowledge Landscape, Cartes de compétences ou Cartographies des Connaissances « métier » sont mises en place pour faire face à la variété et la multiplicité des expertises et connaissances présentes désormais dans les organisations et pour dépasser la difficulté d'accéder à ces expertises via une communication informelle. Elles sont également déployées pour mettre en valeur les connaissances dites « critiques » de l'entreprise. La finalité première de ces cartographies est l'identification de connaissances. La cartographie des connaissances permet de visualiser les domaines de connaissances de l'entreprise dans un cadre compréhensible. En ce sens ces cartographies s'inscrivent dans une des premières étapes spécifiques de la valorisation des connaissances des entreprises: identifier le patrimoine de connaissances et tout particulièrement les domaines de connaissances sur lesquels il faudra faire porter en priorité les démarches de Gestion des Connaissances.

4. Technologies linguistiques

La manipulation des documents textuels [5] pour l'extraction de connaissances qu'il contient est une pratique dont l'importance est reconnue depuis longtemps.

Ces systèmes de traitement automatique prennent en entrée des textes ou ensembles de textes qu'ils transforment pour obtenir en sortie une ou plusieurs représentations du sens. La tâche essentielle de l'opération de transformation consiste à traduire des documents potentiellement ambigus en représentations non ambiguës.

La question de la « compréhension » d'un document textuel, qui est au cœur de toute tâche du traitement automatique de la langue (TAL), renvoie donc à deux problèmes majeurs : le premier concerne la représentation du sens du texte et le second la prise en compte du monde de connaissance de référence. Un système de TAL peut donc commencer l'analyse au niveau du mot pour en déterminer la nature et la structure morphologique, continuer au niveau de la phrase pour déterminer l'ordre des mots, la structure syntaxique et le sens de la phrase entière, avant de s'intéresser enfin au contexte et à l'environnement ou au domaine de référence. Un mot ou une phrase peut avoir un sens spécifique ou une connotation particulière en fonction d'un contexte ou d'un domaine et peut être en résonance avec d'autres mots ou d'autres phrases dans un contexte donné ou en fonction d'un usage particulier.

Pour effectuer une tâche de TAL, on distingue classiquement (pour la langue écrite) six niveaux de traitement :

- le niveau de la segmentation en mots et en phrases ;
- le niveau morphologique qui traite de la manière dont sont constituées les unités lexicales (flexion, dérivation, composition, etc.) et vise à déterminer la catégorie de discours de l'unité considérée ;
- le niveau syntaxique qui détermine la structure des phrases en fonction de la grammaire de référence;
- le niveau sémantique qui traite du sens des mots et des phrases ;
- le niveau du discours qui vise à identifier la structure discursive et argumentative du document;
- le niveau pragmatique qui traite du monde de connaissance de référence, c'est-à-dire qui prend en compte les informations extralinguistiques qui peuvent contribuer à la compréhension du texte.

Cette décomposition en six niveaux est bien sûr toute théorique. Elle ne correspond pas nécessairement au mode de fonctionnement réel de tous les logiciels de TAL. Certains groupent les niveaux 2, 3 et 4 en une seule étape du traitement, alors que d'autres ne prennent pas en compte certaines des étapes mentionnées (par exemple, le niveau pragmatique est rarement pris en compte en tant que tel mais des connaissances de nature pragmatique peuvent être intégrées dans les dictionnaires de référence, en particulier les connaissances métiers). Enfin, les algorithmes utilisés

pour les différents niveaux d'analyse ne procèdent pas tous de la même manière (analyse descendante ou montante, avec ou sans retour arrière, etc.).

4.1. La dimension sémantique dans le TAL

La description sémantique des lemmes s'avère une tâche extrêmement difficile et coûteuse. Les systèmes linguistiques intégrant le niveau d'analyse sémantique sont désormais opérationnels. D'un point de vue fonctionnel, l'apport de la sémantique permet de désambiguïser les textes qui sont analysés. Du point de vue de l'utilisateur, la décision de recourir à ces approches dépend de plusieurs critères :

- la délimitation conceptuelle du domaine : plus le domaine est spécialisé, bien délimité, meilleurs sont les résultats ;
- l'évolutivité du domaine : plus le domaine est stable, moins le système de représentation sémantique devra évoluer, moins la maintenance sera fastidieuse ;

5. Conclusion

Dans ce premier chapitre, il était nécessaire de donner une vision globale sur les documents textuels et les différentes opérations qui peuvent leur être appliqués. Dans les chapitres suivants, nous allons entrer dans le vif du sujet en parlant de la cartographie d'informations (connaissances).

Chapitre II :
La cartographie,
généralités et
méthodes

1. Introduction

L'information textuelle prend de plus en plus d'importance dans l'activité quotidienne des chercheurs et des entreprises. Même si cela l'était déjà auparavant la quantité pose des problèmes d'accès et de recherche d'information. La question se pose de structurer une base documentaire et d'exploiter cette structuration avec un maximum de rendement. L'acquisition des connaissances est un moyen privilégié pour tenter de localiser les problèmes d'extraction de connaissances. Mais l'étude des textes requiert une bonne compréhension des mécanismes linguistiques que l'on peut identifier et analyser automatiquement: il s'agit du traitement automatique du langage naturel. La maîtrise d'un espace est une problématique universelle qui préoccupe l'homme depuis son origine. La cartographie résulte de cette problématique. Cartographier un espace permet de le visualiser en une carte pour mieux l'appréhender. En effet, comprendre quelque chose se dit « voir ». De même, on dit qu' « une image vaut dix milles mots » ou bien encore qu' « une vue d'ensemble est nécessaire ». La métaphore courante entre les processus de la pensée et la vision témoigne du lien étroit qu'il existe entre ce que l'on voit et ce que l'on pense. Dans ce chapitre, nous allons commencer par définir les notions de connaissance, gestion et ingénierie des connaissances, puis introduire la cartographie des connaissances et les différents modèles et techniques pour la cartographie sémantique des documents textuels.

2. La connaissance

La connaissance est⁴ : « Opération par laquelle l'esprit humain procède à l'analyse d'un objet, d'une réalité et en définit la nature : Connaissance intuitive ».

Dans une organisation, on trouve deux grands types de connaissances [7]:

- „ Connaissances implicites: La connaissance peut être innée donc détenue par les employés, les experts, ... et non transcrite sur un support. La connaissance implicite est difficile à formaliser à extraire et à diffuser.
- „ Connaissances explicites: Les connaissances explicites sont transférables physiquement car elles se présentent sous forme de documents.

⁴ Source : la rousse (www.larousse.fr).

En plus la connaissance peut être:

- Interne: L'ensemble des compétences et de l'expérience professionnelle est une importante source interne pour l'entreprise. Elle a l'occasion de renforcer les bases. Pour tout problème, il faut toujours commencer par chercher la solution en interne car elle y est sûrement.
- Externe: L'entreprise doit aussi s'orienter vers l'extérieur pour aller chercher des informations nécessaires au bon développement. Nouveaux produits, nouvelles opportunités, nouvelles connaissances, nouvelles technologies, nouveaux concurrents, ... L'acquisition du savoir externe est indispensable.
- Individuelle: La connaissance est le fait d'individus. Une communauté crée un environnement favorable à la création de connaissances. L'employé dans l'entreprise développe son propre savoir.
- Collective: C'est la connaissance dans un certain contexte d'interaction. Ce n'est pas la somme des connaissances individuelles. L'acquisition du savoir se fait à travers la formation, des routines. L'employé décide de partager le savoir avec les autres.

3. Gestion des connaissances

Une gestion des connaissances [8] a longtemps existé mais de manière non formalisée. C'est grâce aux progrès réalisés en informatique à partir de constats réalisés dans les années 1980 que la gestion des connaissances a émergé. Les entreprises se trouvent alors face à une nouvelle problématique : comment gérer les informations nécessaires à l'activité ? Comment les valoriser ? Comment déterminer les informations les plus pertinentes et extraire de nouvelles connaissances ? L'informatique, qui a prouvé son potentiel durant la deuxième guerre mondiale et pendant la « guerre froide », est mise à contribution et prend alors l'essor qu'on lui connaît. Simultanément, les entreprises découvrent de nouveaux besoins et développent rapidement de nouvelles activités : veille stratégique, gestion de l'innovation, capitalisation des savoirs, gestion du capital intellectuel et humain, intelligence économique, etc. La conséquence directe est un accroissement permanent de la quantité d'information à manipuler. Toutes ces activités sont regroupées sous le nom de *gestion des connaissances* :

Elle consiste à capturer et à représenter les connaissances des organisations pour faciliter leur accès, leur partage et leur réutilisation.

4. Ingénierie des connaissances

L'Ingénierie des connaissances [9] constitue depuis de nombreuses années un domaine actif des recherches menées en intelligence artificielle autour de la conception et de la réalisation des systèmes à base de connaissances (SBC). À l'instar de bien d'autres disciplines modélisatrices, elle consiste à concevoir des systèmes dont le fonctionnement permet d'opérationnaliser des connaissances portant sur le traitement ou la résolution d'un problème donné. La résolution (semi-)automatique de problèmes implique deux étapes essentielles : la modélisation du problème et d'une méthode de résolution dans un cadre théorique donné, l'opérationnalisation informatique du modèle obtenu. Longtemps, le cadre théorique de l'Ingénierie des connaissances fut celui de l'acquisition des connaissances : modélisation psychologique ou empirique des connaissances d'un expert dans le but de les coder dans un système expert (SE). La période actuelle se concentre davantage sur la modélisation conceptuelle du monde : on tente, le plus souvent à partir d'une formulation linguistique du problème, transcriptions d'interviews d'experts, descriptions techniques, notices de maintenance, etc., d'élaborer une représentation qualitative et formelle du problème.

5. La cartographie sémantique

La cartographie sémantique [10] permet de représenter les informations en les liant grâce à leur sens (sémantique = qui a rapport avec le sens).

Selon le comité Français de cartographie⁵ : « La cartographie est l'ensemble des études et des opérations scientifiques, artistiques et techniques, intervenant à partir des résultats d'opérations directes ou de l'exploitation d'une documentation, en vue de l'élaboration et de l'établissement de cartes, plans et autres modes d'expression, ainsi que dans leur utilisation. »

⁵ <http://www.lecfc.freesurf.fr>

6. Procédé d'une cartographie d'information

Une simple vision [11] d'une cartographie d'information est la suivante : Au départ des données brutes (pas encore manipulées) sont collectées généralement grâce à l'aide d'un procédé automatisé. L'utilisateur extrait un sous-ensemble de données intéressantes organisées d'une manière plus structurée. Cette forme plus structurée peut alors être associée à une représentation visuelle par association des propriétés des données aux attributs visuels. Finalement, la représentation visuelle peut-être manipulée de manière interactive par l'utilisateur en obtenant différentes vues de la même information.

7. Objectifs de la cartographie d'informations

Il s'agit de fournir à l'utilisateur une compréhension qualitative du contenu de l'information. En effet, les apports de la cartographie peuvent être résumés ainsi [12, 11]:

- La cartographie est tout un processus délicat au bout du quel, une grande d'information textuelle est nettoyée, résumé et représentée sous forme visuelle ;
- Cette visualisation doit permettre à l'utilisateur final de faire des découvertes, proposer des explications ou prendre des décisions ;
- Ces actions peuvent se faire aussi bien sur des motifs (clusters, tendances, émergences, anomalies) ou sur des ensembles d'éléments ou encore sur des éléments isolés ;
- Communiquer efficacement des informations au travers d'une représentation graphique via des cartes cognitives⁶ ;
- Faciliter la découverte de connaissances grâce à une représentation graphique issue de l'analyse d'un corpus d'informations via des cartes sémantiques ;
- Les visualisations graphiques ont le pouvoir de permettre à l'homme de manipuler des structures bien plus complexes représentées par une visualisation (représentations externes) que dans la mémoire de travail visuelle et verbale (représentations internes) ;

⁶ « Une carte cognitive est un modèle conçu pour représenter la façon dont une personne définit un problème particulier... Toutefois ce n'est pas un modèle générale de la façon de penser de quelqu'un [sic !] » (EDEN, « a response to Watson DeSanctis and Poole », 1988)

- Elles permettent aussi de percevoir l'émergence de propriétés dans les données cartographiées qui ne sont pas anticipées ;
- Les visualisations permettent de mettre en évidence des problèmes dans les données, dans leur collecte. Avec une visualisation appropriée, les erreurs dans les données sont rapidement perceptibles ;
- La visualisation permet de percevoir simultanément des propriétés à grande et à petite échelle sur les données ;
- La visualisation facilite la formation d'hypothèses sur les données.

8. Approches de cartographie de connaissances

Réaliser une cartographie des connaissances consiste à faire un inventaire détaillé des connaissances détenues par les membres d'une organisation et jugées utiles pour assurer sa performance. On distingue deux approches [13]:

8.1. Approche par le processus

L'approche par les processus vise à déterminer les connaissances critiques dont la perte entraînerait des dysfonctionnements. On distingue deux types de cartes associées à cette approche :

- *Les Knowledge source maps*: qui sont des cartographies d'expertise. Pour établir ce type de carte, on peut utiliser la démarche GAMETH , élaborée par Michel Grundstein en 2002. Elle vise à repérer, localiser et caractériser les connaissances critiques d'une organisation. La connaissance repose sur deux catégories le savoir-être et le savoir-faire. Le savoir-être regroupe les composantes qui influent la façon dont un individu ou un groupe d'individu construit une vision du monde qui lui est propre. Le savoir-faire regroupe les composantes qui permettent à un individu ou à un groupe d'individus de résoudre un problème.
- *Les knowledge application maps*: servent à recenser les connaissances liées à un problème ou à un besoin spécifique. Tseng et Huang proposent en 2005 une démarche pour identifier les connaissances cruciales d'une organisation (c'est à dire Les connaissances nécessaires pour résoudre les problèmes portant sur un objectif donné, et qui devraient être capitalisées). Leur approche est basée sur une analyse quantitative de l'information recueillie en interviewant des experts. Tseng et Huang proposent une procédure

algorithmique à partir des données recueillies afin de déterminer quatre ensembles qui caractérisent l'importance de la connaissance:

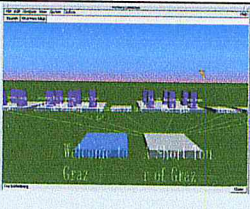
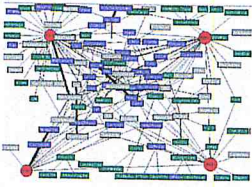
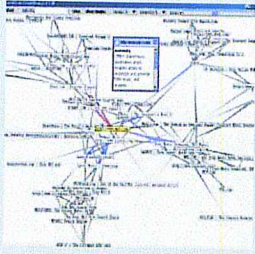
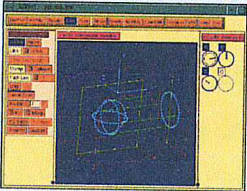
- Les connaissances vitales : ce type de connaissances est considéré par les auteurs comme très important. Elles doivent être répertoriées et situées dans leur contexte.
- Les connaissances devant être acquises rapidement : ce type de connaissances est considéré par les auteurs comme important car il permet de résoudre facilement certains types de problèmes fréquents.
- Les connaissances dites saisonnières : ce type de connaissances n'est pas utile lors de la résolution de la plupart des problèmes.
- Les connaissances insignifiantes : ce type de connaissances n'est pas collecté et aucune action n'est recommandée, sauf en cas de besoin particulier.

8.2. Approche par les domaines

La cartographie des connaissances « domaine » d'une organisation est associée à des pratiques métier au sens de la théorie de communautés de pratique de Wenger (1998). Elle se réalise à partir l'analyses de documents ou de témoignages d'acteurs. On distingue deux principaux types de cartes associés à cette approche.

- *Les knowledge source maps* : Elles représentent des cartographies de stock de connaissances. La méthode des arbres de connaissances (Authier & Levy, 1992) permet de les construire. Un arbre de connaissances représente l'ensemble des savoirs que possède une communauté, que celle-ci soit une entreprise, un quartier, un organisme de formation, une région, etc. La compétence ou l'objet de connaissance ainsi représenté peut être l'attribut d'une ou de plusieurs personnes. Dans le tronc de l'arbre se trouvent les savoirs de base, ceux que les membres de la communauté ont acquis en premier, tandis que les feuilles correspondent aux savoirs spécialisés; les branches, elles, rassemblent les savoirs qui se trouvent toujours associés dans certains blasons (liste des compétences ou des objets de connaissances détenus par une même personne). Le blason constitue le cœur du système ; propre à chaque individu, c'est une représentation graphique de ses savoirs et savoir-faire, y compris ceux qui sont nés de l'expérience des individus appartenant à un groupe.

Chapitre II: Cartographie, généralités et méthodes

		Navigator, Information Pyramids, GopherVR		
réseaux	Réseaux de personnes 	NetMap, inFlow, ContactMap, TheBrain, SemNet, Harmony Locap Map et Harmony Locap Map 3D	Grande interactivité	The Brain ne permet pas de visiter les niveaux bas de l'hierarchie.
	Réseaux de documents 	Butterfly, CiteSpace II Mapping Topic Burst TouchGraph, CiteWiz.	Bonne apprehension du domaine scientifique	Illisibilité des réseaux volumineux
	Plateformes génériques	graphViz, Jung, GUESS, Piccolo,	Grande souplesse de paramétrage, visualisation à la carte	Non destinés aux novice, il faut toucher au code source.
Multidimensionnelle	Nuage de points 	Xgobi, Envision, Cognos Visualizer, SpotFire, FilmFinder, Miner3D.	Filtrage des informations à travers des requêtes	Difficulté de lecture,
	Les diagrammes	Attribute	Visualisation	Une période

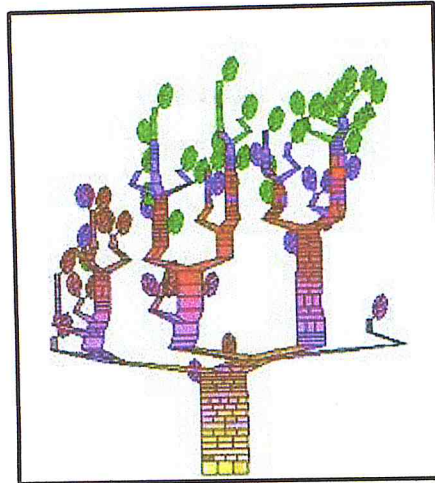


Figure 1: Arbre de connaissances

- *Les knowledge asset maps* : Elles permettent de recenser les besoins en connaissance pour un domaine particulier. Pomian et Roche proposent en 2002 une méthode de cartographie des connaissances domaines d'une organisation en se basant sur la distinction entre les connaissances tacites et les connaissances explicites.
 - Les connaissances tacites sont identifiées au cours d'entretien d'experts appartenant à l'organisation puis classées selon leur degré d'utilité et leur probabilité à être réutilisées.
 - Les connaissances explicites sont identifiées à partir de l'analyse des documents émis par l'organisation. Chaque document est classé selon quatre critères : le niveau de lisibilité, de clarté, de pertinence et d'accessibilité.

Une autre méthode permet de réaliser des Knowledge asset maps, la méthode M3C (Toukara & al, 2005). Cette méthode a pour objectif de représenter de façon hiérarchique les différents domaines de connaissances d'une organisation. La mise en œuvre de la méthode M3C nécessite de suivre plusieurs étapes de façon chronologique :

- L'identification des connaissances de l'organisation : cette identification est réalisée à partir des différents documents émis par l'organisation (statuts, organigramme, descriptions des stratégies, études, publication de résultats etc.) ;

Dans les faits, le recours aux technologies de visualisation de l'information peut poursuivre deux objectifs distincts :

- communiquer efficacement des informations au travers d'une représentation graphique ;
- faciliter la découverte de connaissances en utilisant une représentation graphique issue de l'analyse d'un corpus d'informations.

A travers les années, plusieurs méthodes et outils ont vu le jour. Mais leur comparaison ne peut pas être faite de manière objective. En effet, la diversité qui prévaut dans ce domaine est due à la diversité du type d'applications finales et de ce fait, le choix d'un modèle de visualisation revient au choix de l'approche désirée (par processus ou par domaines).

11. Conclusion

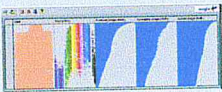
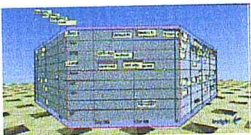
La cartographie a pour finalité de visualiser (sous forme de carte) un ensemble de données abstraites ou scientifiques son objectif est de transmettre une information. Au fil des siècles, la cartographie est devenue un art d'expression mais aussi, un outil d'analyse et de communication.

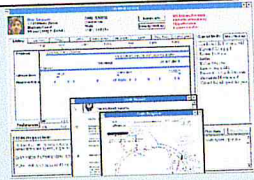
Le nombre de paradigmes de cartographie est très important, par conséquent il est possible de construire un nombre quasiment infini de cartes différentes d'un même espace de données. Dans un tel contexte, comment construire une carte qui aidera effectivement ses utilisateurs à appréhender leur espace informationnel ? Dans ce qui suit, nous allons dans un premier lieu, affiner le domaine d'application en introduisant les spécificités des notices pharmaceutiques. Ce qui nous permettra de dégager les méthodes de cartographies qui nous conviennent afin de pouvoir construire notre propre méthode (originale et personnalisée) par la suite.

- La construction d'une première version de cartographie : Cette première version est conçue en suivant un processus itératif : appropriation du problème, co-construction de la représentation et validation ;
- Entretiens individuels et collectifs d'experts : Ces entretiens ont pour objet de discuter, modifier ou valider la première version de la cartographie ;
- Élaboration de critères de classification de domaines de connaissance : Cette étape permet de classer chaque domaine de connaissance de la carte réalisée précédemment en fonction de son caractère plus ou moins critique pour l'organisation ;
- Réalisation de la carte de connaissance : La version finale de la cartographie se fait après échantillonnage, collection et scoring des données.

9. Types de cartographie et leurs outils

Dans le tableau qui suit, nous allons synthétiser une étude faite par Solveig Vidal en 2006 [11] concernant quelques types de visualisation et les outils qui les implémentent :

Type	Approche	Outils	Avantages	Inconvénients
Linéaire	Les gros tableurs 	Tablens, Devis, SeeSoft.	Vue d'ensemble, filtrer détailler et trier les données et y détecter les anomalies et les tendances.	Données d'intérêt non exportable
	Les murs fuyants 	TimeWall, Perspective Wall	Suivi temporel des événements, vue d'ensemble sur les données.	Difficulté pour naviguer dans le timeline
	Les frises chronologiques	lifeStream, LifeLine	//	

				
hiérarchique	Liste indentée	Microsoft Windows Explorer	Simplicité de compréhension.	Mauvaise navigation,
	Diagramme	Space Tree, Cheops, Syn Vis Magic Eye	Optimisée par rapport aux listes indentées, affichage de l'hierarchie totale.	Mauvaise visibilité
	Surfacique	Market map, Treemap	Arbre de plusieurs dizaines de milliers de nœuds.	Statique sans interaction
	Géométrique hyperbolique	Star Tree, H3 3D, Hyperbolic, Browser, Walrus.	Meilleure visualisation des nœuds en périphérie	Utilisation nécessite entraînement.
	Canonique (3D)	Cat-a-cone, LyberWorld	Vue en 3D, excellente ergonomie.	Distinction difficile entre les cônes
	Paysage d'informations	Harmony Information Landscape, File System	Bonne visualisation et survol dynamique	Grande consommation de la mémoire

Chapitre III :
Spécificités des
notices
pharmaceutiques

1. Introduction

Au cours de ce chapitre, nous allons aborder les différentes sortes de médicaments gérés et dispensés par les médecins et en pharmacie. Ceci nous aidera à mieux les comprendre et du coup à savoir comment les traiter, gérer, et représenter automatiquement.

2. Les notices pharmaceutiques

2.1. Qu'est-ce qu'une notice ?

La notice contenue dans la boîte du médicament [14] est une information destinée au patient. Mode d'emploi, elle consigne également des informations clés pour aider chacun à tirer le maximum de bénéfices de son traitement dans les meilleures conditions de sécurité. Pour chaque spécialité pharmaceutique, la notice contient obligatoirement les rubriques suivantes :

- Composition ;
- Indications thérapeutiques ;
- Enumération des informations nécessaires avant la prise du médicament ;
- Instructions nécessaires au bon usage ;
- Description des effets indésirables ;
- Conditions de conservation.

Si elle signale quels sont les bénéfices du traitement, elle aborde aussi clairement les risques qui y sont liés et les moyens de les éviter ou de les minimiser. Des informations d'autant plus importantes à consulter que le médicament est pris en automédication. Mais, si le doute persiste après leur lecture, il est essentiel de prendre conseil auprès de son médecin ou de son pharmacien.

2.1.1. Des informations toujours plus accessibles

Longtemps jugées trop compliquées par les patients, les notices font désormais d'une attention particulière de la part des laboratoires, qui ne se contentent plus d'y faire figurer le maximum d'informations obligatoires mais qui veillent aussi à leur clarté. Et pour cause, l'enjeu est de taille : de leur compréhension dépendent l'efficacité du traitement et la sécurité des patients. En 1988, cette volonté de clarté a conduit à élaborer, au niveau européen, des

recommandations en direction des laboratoires pour que la notice soit rédigée en des termes compréhensibles et bénéficie d'une présentation lisible. A titre d'exemple sont conseillés :

- L'utilisation de caractères de taille et de police faciles à lire ;
- La rédaction de paragraphes et de phrases courtes (moins de 20 mots) ;
- Le recours à des interlignes et des espacements suffisants ;
- Un bon contraste lors de l'impression...

Depuis 2004, les entreprises du médicament se livrent même à des tests de lisibilité auprès de groupes « cibles » de patients. L'objectif est de s'assurer que les patients sont capables, d'une part de trouver l'information (lisibilité), d'autre part de la comprendre (clarté) et enfin de la mettre en application (facilité d'utilisation).

Depuis 2009, pour assurer l'accès de l'information aux malvoyants, les industriels doivent faire figurer sur l'emballage de tous les médicaments le nom et le dosage en braille et fournir, sur demande, des notices adaptées. Une obligation qui avait été largement anticipée par les laboratoires pour les médicaments les plus utilisés.

2.2.Format d'une notice

Les notices pharmaceutiques doivent comporter toujours plus d'informations qu'il faut ajuster à un format lisible – nécessitant plus de surface de papier, offrant une lisibilité parfaite, et plié dans un emballage le plus petit possible. Pour avoir une vision claire des perspectives qui peuvent être considérées dans l'optique de notre projet, nous allons définir les notions rudimentaires [15] concernant les médicaments.

3. Généralité sur la pharmacologie

3.1.Qu'est-ce qu'un médicament

On entend par médicament toute substance ou composé présenté comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que tout produit pouvant être administré à l'hôpital ou à l'animal en vue d'établir un diagnostic ou restaurer, corriger ou modifier leurs fonctions organiques.

Chapitre III : Spécificités des notices pharmaceutiques

Un médicament agit par l'intermédiaire d'un ou plusieurs constituants appelés *principes actifs* (substances réellement actives), qui sont associées à des *excipients* (substances non actives qui permettent la préparation et l'administration du médicament).

Le code de la santé définit les spécialités pharmaceutiques comme « tout médicament préparé à l'avance, présenté sous un conditionnement particulier et caractérisé par une dénomination spéciale. Aucune spécialité ne peut être débitée à titre gracieux ou onéreux si elle n'a pas reçu auparavant une Autorisation de Mise sur le Marché délivrée par le Ministère de la Santé.

3.2.Médicaments essentiels

Ce sont des médicaments qui satisfont aux besoins fondamentaux de la majorité des populations en matière de soins de santé. Ce sont des médicaments pour lesquels il existe des données sûres et suffisantes sur l'efficacité et les effets secondaires, et qui ont un moindre coût. Ils doivent être disponibles à tout moment.

Une liste de médicaments essentiels a été établie par le Ministère de la santé, visant à la prise en charge efficiente des pathologies les plus courantes : *soigner le maximum de maladies (95%) avec le minimum de médicaments.*

3.3.Spécialités et génériques

Tout médicament découvert ou synthétisé par un laboratoire pharmaceutique est la propriété de celui-ci. Cette propriété est protégée par un brevet qui confère le monopole d'exploitation pendant une vingtaine d'année. Le laboratoire donne au médicament un nom de fantaisie ou nom commercial et son conditionnement est particulier. On parle alors de *spécialité*.

Au moment où le brevet d'exploitation expire, tout laboratoire peut produire ce médicament. Certains laboratoires produisent alors des médicaments *génériques*, désignés par leur Dénomination Commune Internationale (DCI), qui fait référence au principe actif et est la même dans tous les pays du monde. Il faut remarquer que les spécialités portent aussi un nom DCI qui figure obligatoirement en dessous du nom commercial. Ainsi, un même médicament a un seul nom DCI, mais peut avoir plusieurs noms commerciaux.

3.4. Les familles de médicaments

Tous les médicaments n'agissent pas de la même manière, et ils ne traitent pas tous les mêmes maladies. C'est pour cela qu'ils appartiennent à des *familles de médicaments* ou *classes thérapeutiques* que nous allons tenter de résumer dans le tableau suivant [16] :

Famille	Caractéristiques	Médicaments	Conseils
Antalgiques (Analgésiques)	Diminuent la douleur (maux de tête, états fébriles, douleurs dentaires)	Acide Acétylsalicylique (Aspirine), Acétylsalicylate de lysine injectable, Ibuprofène, Paracétamol	Ne pas délivrer l'Acide Acétylsalicylique et l'ibuprofène à un patient souffrant de gastrite ou de brûlures d'estomac, un enfant ou une femme enceinte.
Antipyrétiques	Abaisser la fièvre	//	-
Anti-inflammatoires	Atténuer les rougeurs et les douleurs liées à des inflammations mais ne les illuminent pas	l'Acide Acétylsalicylique (Aspirine), Ibuprofène, Diclofénac.	-
Anti-infectieux	Préconisés pour le traitement des infections telles que les microbes (bactéries, virus, parasites, champignons)	-	-
Antibiotiques	Arrêtent la multiplication des bactéries	Pénicillines, tétracyclines, Chloramphénicol, Gentamicine, Cotrimoxazole, Érythromycine.	<ul style="list-style-type: none"> • Ne pas l'employer à tort et à travers ; • Respecter la posologie et

Chapitre III : Spécificités des notices pharmaceutiques

			la durée de traitement.
Antiparasitaires	Tuent les parasites internes et externes	-	-
Antipaludéens	Agissent contre le paludisme (du aux piqûres de moustiques)	Chloroquine, sulfadoxine-pyriméthamine, sels de quinine	-
Anti-amibiens	Agissent contre l'amibe (douleurs abdominales, selles glaireuses)	métronidazole	Eviter la consommation d'alcool durant le traitement.
Anti-bilharziens	Contamination parasitaires lors de baignades dans les eaux stagnantes	praziquantel	-
Anthelminthiques	Vers intestinaux qui provoquent un amaigrissement et une altération de la santé	Nicosamides, mébendazole	-
Antifongiques (antimycosiques)	Traitements des champignons de type mycosique (cutané, vaginal, digestif,...)	Nystatine, acide salicylique, acide benzoïque.	-
Anti-diarrhéiques	Forte déshydratation	Métronidazole, cotrimoxazole, amoxicilline	Réhydratation orale pour restituer l'eau et les sels minéraux
Anti-spasmodiques	Diminuent les spasmes gastro-intestinaux et génito-urinaires	Atropine, butylscopolamine, butylhyoscine	-

Antiémétiques	Traitent les vomissements et les nausées	Métoclopramide, métopimazine, chlorpromazine	Administration par voie orale ou par injection.
Antiépileptiques (anticonvulsivants)	Traitement des convulsions accompagnées (ou non) de perte de connaissance	Phénobarbital, diazépam	-
Antiallergiques	Traitements des réactions allergiques (démangeaisons, boutons, éternuements)	Chlorphéniramine, prométhazine	Les infections allergiques peuvent aussi être traitées par des anti-inflammatoires
Médicaments de l'appareil respiratoire	Asthme, crises d'étouffement,	Aminophylline, salbutamol, corticoïdes	Les infections respiratoires peuvent aussi être traitées par des antibiotiques
Médicaments ophtalmologiques	Traitements des différentes maladies des yeux (conjonctivite,	Nitrate d'argent collyre, chloramphénicol	Ces manifestations peuvent être traitées par des antiseptiques ou des antibiotiques. Les flacons de collyres ne doivent pas être utilisés plus de 15 jours.

Tableau 2: Familles des médicaments

4. Conclusion

Dans ce chapitre, il était nécessaire de donner une vision globale sur les notices pharmaceutiques et toutes les notions qui se rapportent au domaine pharmaceutique. Cette étape était très conciuante pour nous telle qu'elle nous a permis de dégager les axes sur les quels nous pouvons intervenir pour garantir une activité décisionnelle basée sur la cartographie sémantique et la

Chapitre III : Spécificités des notices pharmaceutiques

classification intelligente des notices pharmaceutiques. Dans ce qui suit, nous allons présenter les différents constats et les justifications quant à l'approche que nous proposons.

Chapitre IV :
Analyse des
besoins et
modélisation du
systeme

1. Introduction

Après avoir étudié les documents textuels électroniques et vu les opérations qu'on peut leur affecter en terme de traitement linguistique, représentation et gestion, nous avons consacré un chapitre abordant les spécificités des notices pharmaceutiques (choisies, pour leur ambiguïté, comme support de test) pour savoir quel type de traitement pourrait leur être appliqué afin de capitaliser le contenu sémantique qu'elles contiennent. Finalement, le parcours bibliographique nous a permis de dégager les grandes lignes nécessaires à cerner notre problématique de départ, et de ce fait proposer une solution à la prise en compte -dans une optique décisionnelle- de la dimension sémantique contenue dans les notices pharmaceutiques par le biais de visualisations non seulement graphiques mais aussi formelles. Dans ce chapitre, nous allons justifier notre approche et présenter la modélisation du système proposé.

2. Constat

L'étude menée sur les notices pharmaceutiques nous a permis de constater les points suivants :

- Une notice est un document textuel structuré en section,
- Une notice contient l'information nécessaire à la compréhension des composants, des indications et contre-indications, et de la posologie et du mode d'administration du médicament. Cependant cette information, malgré le fait qu'elle ait subi des améliorations visant à la simplifier (comme nous l'avons vu dans le 3^{ème} chapitre), elle reste parfois difficile à comprendre,
- Le volume textuel condensé dans une notice peut être frustrant et bloquant quant à une activité décisionnelle urgente (nous allons revenir sur ce point plus loin dans le chapitre),
- Les médicaments peuvent être classifiés en familles et la DCI peut être exploitée pour une éventuelle classification automatique,
- Dans une optique décisionnelle urgente, certaines sections des notices pharmaceutiques peuvent être volontairement omises.

A la lumière de ces constats, nous avons imaginé une situation décisionnelle urgente qu'une cartographie sémantique et une classification automatique des notices pharmaceutique pourrait régler. La situation est la suivante : on suppose une pharmacie, où le vendeur n'est pas un pharmacien de formation (i.e. non expert du domaine pharmaceutique). L'étiq ue voudrait que le pharmacien en chef de la pharmacie lui enseigne les notions rudimentaires relatives au domaine pharmaceutique mais en réalité, les choses ne se passent pas toujours ainsi. Supposons maintenant que ce vendeur soit impliqué dans une situation où il doit vendre un médicament sans ordonnance à un patient et compte tenu de la responsabilité qu'incombe cette tâche, il ne faut pas que le vendeur fasse une erreur qui coûterait peut être la vie au patient. Notre solution dans ce cas est de doter la pharmacie d'un logiciel de capitalisation des connaissances contenues dans les notices pharmaceutiques en les représentant sous forme graphique, qui aide la compréhension cognitive rapide, et en classant les notices selon leurs familles afin de faciliter leur recherche.

3. Processus global d'une cartographie sémantique

Pour pouvoir présenter d'une manière plus claire le fonctionnement de notre système, nous nous devons d'exposer le processus global d'une cartographie d'information. Pour ce faire, nous allons adopter une vision très théorique qui décompose l'analyse d'un texte en trois étapes successives à savoir :

- *L'analyse terminologique*, qui est dédiée à l'analyse de la structure des phrases et de la structure des mots ;
- *L'analyse sémantique*, qui s'intéresse au sens des phrases considérées individuellement ;
- *La transformation de vue*, qui s'attache à remettre un contexte autour des phrases ;

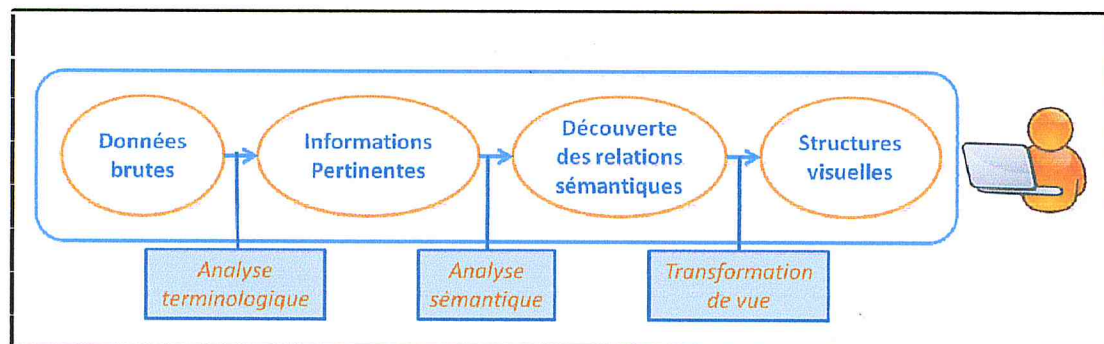


Figure 2: Processus global de cartographie

Donc comme le montre la figure 2, le processus de la cartographie démarre à partir d'un ensemble de données brutes. On entend par données brutes tout le contenu textuel de la notice qui à un stade primaire de la cartographie, n'a subi aucun traitement (on parle de nettoyage des données). A cet ensemble de données brutes nous allons effectuer une analyse terminologique, composée à son tour de deux phases. La première étant syntaxique et la seconde lexicale. Au bout des ces deux phases nous obtenons un ensemble restreint de données dites pertinentes qui va faire l'objet d'une analyse sémantique. Cette analyse sémantique va nous permettre de prélever les éléments susceptibles de nous aider pour classer la notice d'une part et de découvrir les relations sémantiques entre les différents mots pertinents afin de construire la vue graphique. Enfin, ces informations de type relationnelles vont nous permettre de monter à un niveau structurel d'analyse au bout duquel la représentation visuelle de la notice s'affichera à l'utilisateur qui selon son degré de satisfaction décidera de sauvegarder ou non la cartographie. Si l'utilisateur décide de la sauvegarder alors la notice va être classée automatiquement parmi les autres notices du corpus.

Ceci étant une vision assez globale du fonctionnement de notre système final, nous allons dans ce qui suit détailler la modélisation de chacune des étapes du processus.

3.1. Analyse terminologique

Afin d'effectuer cette phase, nous nous sommes inspiré des techniques de nettoyage de données dans le traitement automatique de la langue TAL qui s'axent généralement sur deux étapes essentielles :

- *Analyse syntaxique* : qui a pour objectif d'identifier les mots du texte (simples, composés, noms propres, abréviations) et leurs traits (genre et nombre). L'analyse lexicale permet de nettoyer les données de départ en :
 - Homogénéisant le texte à une même forme (généralement minuscule) ;
 - Éliminant les mots non pertinents dits *vides*⁷ (comme les chiffres, les prépositions, les déterminants, les abréviations,...) ;

⁷ Une petite liste exhaustive des mots considérés comme vides dans la langue française peut être trouvée ici : <http://www.ranks.nl/stopwords/french.html>

Chapitre IV : Analyse des besoins et modélisation du système

- Une dernière étape dite *normalisation* s'impose généralement dans la phase syntaxique, durant laquelle tous les mots seront réduits à leurs formes canonique (infinitif pour les verbes et singulier pour les noms), on parle de *lemmatisation*. Mais aussi en ne gardant que la racine d'un mot dans le cas où celui-ci possède de nombreuses déclinaisons (exemple : pour les mots compter, compteur, comptage, compte,... uniquement la racine « compt » est gardée), on parle de *radicalisation*. Cette étape permet d'éliminer les doublons et la redondance inutile.

Remarque :

Pour notre cas, nous allons en plus du nettoyage classique des données, éliminer les sections superflues pour une prise de décision par la visualisation graphique (comme la posologie, et la composition, ...). Au fait nous ne garderons de la notice que 3 sections : DCI afin de pouvoir classer la notice, Indications pour savoir dans quels cas le médicament est préconisé, et effets indésirables pour savoir dans quels cas il est interdit d'administrer le médicament à un patient. Aussi nous n'allons pas procéder à l'étape de radicalisation dans de manière effective. En effet, nous allons nous contenter de transformer les mots pluriels au singulier car il est généralement improbable de trouver plusieurs déclinaisons du même mot au sein d'une notice. De plus, pour la construction de la représentation graphique nous auront besoin d'afficher le mot dans sa globalité et non une partie de ce dernier.

- *Analyse lexicale* : L'analyse lexicale a pour objectif d'identifier les mots du texte (simples, composés, noms propres, abréviations) et leurs traits caractéristiques (genre et nombre). L'analyse lexicale se décompose en deux étapes :
 - la segmentation, dont le but est de découper le texte en phrases puis en mots distincts ;
 - l'étiquetage, dont l'objectif est d'identifier *la bonne catégorie syntaxique (verbe, nom...)* des mots selon le contexte.

Remarque :

Pour des contraintes liées au temps et pour absence total, à notre connaissance, d'outils de traitement du langage (français) personnalisés pour le domaine pharmaceutique, et en ne perdant pas de vue notre objectif premier qui consiste en la

cartographie des notices pharmaceutiques, nous avons décidé de ne pas approfondir d'avantage l'analyse lexicale en négligeant l'étape de découverte des catégories syntaxiques et les spécificités liées au lexique (champs lexical, ...) et nous nous sommes contentés de l'étape de segmentation.

La figure 3 synthétise les deux phases de l'analyse terminologique. Où la représentation niveau I correspondrait aux informations pertinentes.

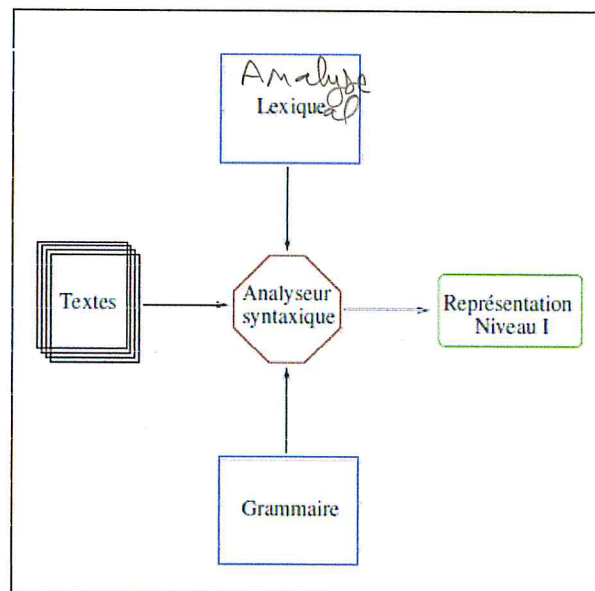


Figure 3: Analyse terminologique

3.2. Analyse sémantique

Au sens littéral [17], la sémantique vise à l'étude du sens hors contexte. Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux de traitements précédents, par conséquent les réalisations qui sont opérationnelles sont peu nombreuses, et elles concernent des applications très **limitées** où l'analyse sémantique se réduit à un domaine parfaitement circonscrit ; par contre, on est encore loin de savoir construire en grandeur réelle des analyseurs sémantiques **généraux** qui couvriraient la totalité de la langue et seraient indépendants d'un domaine d'application particulier.

Dans notre cas, nous entendons par sémantique toutes les informations susceptibles de nous aider à classer la notice dans la famille de médicament qui convient. En l'occurrence et après une étude des spécificités des notices pharmaceutiques, nous avons conclu que la DCI, nous suffisait amplement pour une

telles tâches. Donc le fait d'extraire la DCI, nous permettra de capturer la dimension sémantique permettant de situer le contenu informationnel dans son contexte en classant la notice dans la bonne catégorie.

3.3. Transformation de vue

Le traitement sémantique [17] prend comme unité d'analyse la **phrase**, et conduit à représenter sa partie significative. Ces phrases, dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de **mots** identifiés par l'analyse morphologique (lexicale), et regroupés en **structures** par l'analyse syntaxique. Ces mots et ces structures constituent autant d'**indices** pour le calcul du sens : on pourrait dire, que le sens résulte de la double donnée du sens des mots et du sens des relations entre mots. Dans cette optique, aux trois sections conservées de la notice que nous allons traiter (nettoyer, segmenter, normaliser,...etc) séparément puis on procède à la découverte des relations hiérarchique au sein de chacune d'elles afin de pouvoir construire la vue finale qui sera présentée à l'utilisateur.

Parmi les différents types de représentations (tableau n° 1), nous nous sommes particulièrement intéressés aux représentations hiérarchiques de type diagramme. Au fait nous n'allons pas nous contenter d'adopter une méthode de représentation classique mais plutôt s'inspirer de ce modèle pour proposer une représentation hiérarchique arborescente comme celle proposée par [12] mais la différence réside dans la formalisation de la présentation. En effet, nous allons faire le double pari le fournir une représentation *graphique* donc facilement compréhensible par l'être humain mais aussi *formelle* car sauvegardée selon les préceptes Ontologiques (voir Annexe A) et donc facilement exploitable par la machine. On parle alors de taxonomies de concepts, où les concepts sont les différents termes pertinents pour un domaine donné.

La figure 4 représente la transformation de vue selon [12].

Conception //

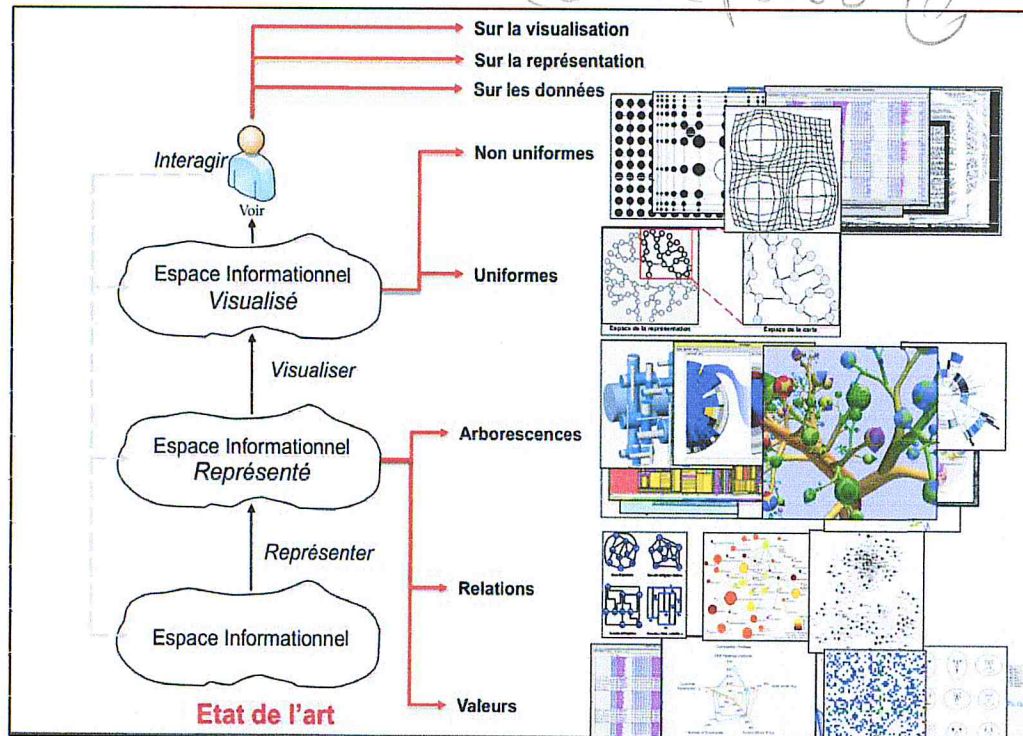


Figure 4: Architecture de transformation de vues

3.3.1. Qu'est-ce qu'un arbre

La structure d'arbre [18] est très utilisée en informatique. Sur le fond on peut considérer un arbre comme une généralisation d'une liste car les listes peuvent être représentées par des arbres. La complexité des algorithmes d'insertion de suppression ou de recherche est généralement plus faible que dans le cas des listes (cas particulier des arbres équilibrés). Les mathématiciens voient les arbres eux-mêmes comme des cas particuliers de graphes non orientés connexes et acycliques, donc contenant des sommets et des arcs (voir figure 5):

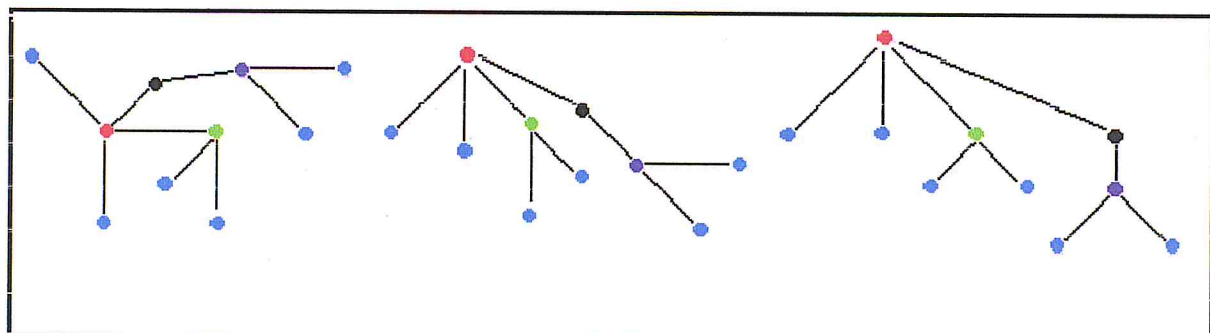


Figure 5: quelques formes arborescentes

3.3.2. Taxonomies

Les taxonomies [19] sont utilisées de façon intensive dans la mise en place d'ontologies, générales ou de spécialité. Les systèmes d'ontologies les plus récents postulent une architecture de représentation de connaissances quelque peu avancée qui nécessite un approfondissement de la notion de taxonomie.

Dans un premier lieu, les nœuds des taxonomies sont des concepts, donc des abstractions sur des connaissances. À chaque nœud, donc concept, est associé, dans une notice donnée, ses différents termes (les différents mots ou groupes de mots utilisés pour désigner ce concept). Souvent, le terme le plus usuel est utilisé pour nommer le nœud. Les différents termes sont autant de termes synonymes, ou presque, compte tenu du niveau de granularité considéré dans la taxonomie.

4. Architecture du système

Dans des systèmes de visualisation complexes (traitant des données de l'ordre de millions d'enregistrements dans le cadre des entrepôts de données⁸ par exemple), l'utilisateur est très sollicité afin de corriger ou modifier les vues présentées. Donc l'aspect interactionnel est un aspect primordial pour la visualisation graphique des données d'un datawarehouse. Dans notre cas, puisque le corpus documentaire est assez petit tel que chaque notice a une granularité de l'ordre de quelques centaines de mots seulement, nous avons opté pour une automatisation totale du processus de cartographie et de classification des notices pharmaceutiques. Ainsi, l'utilisateur n'a qu'à sélectionner la notice, il lance les tâches de traitement de langue, puis de cartographie et enfin de classification une fois qu'il est satisfait de la visualisation. La figure 6, illustre l'architecture d'interaction entre l'utilisateur et notre système de Cartographie Sémantique et de Classification des Notices Pharmaceutique (CSCNPharm).

⁸ Selon www.journaldunet.com, Un datawarehouse (ou entrepôt de données) est un serveur informatique dans lequel est centralisé un volume très important de données consolidées à partir des différentes sources de renseignements d'une entreprise (notamment les bases de données internes).

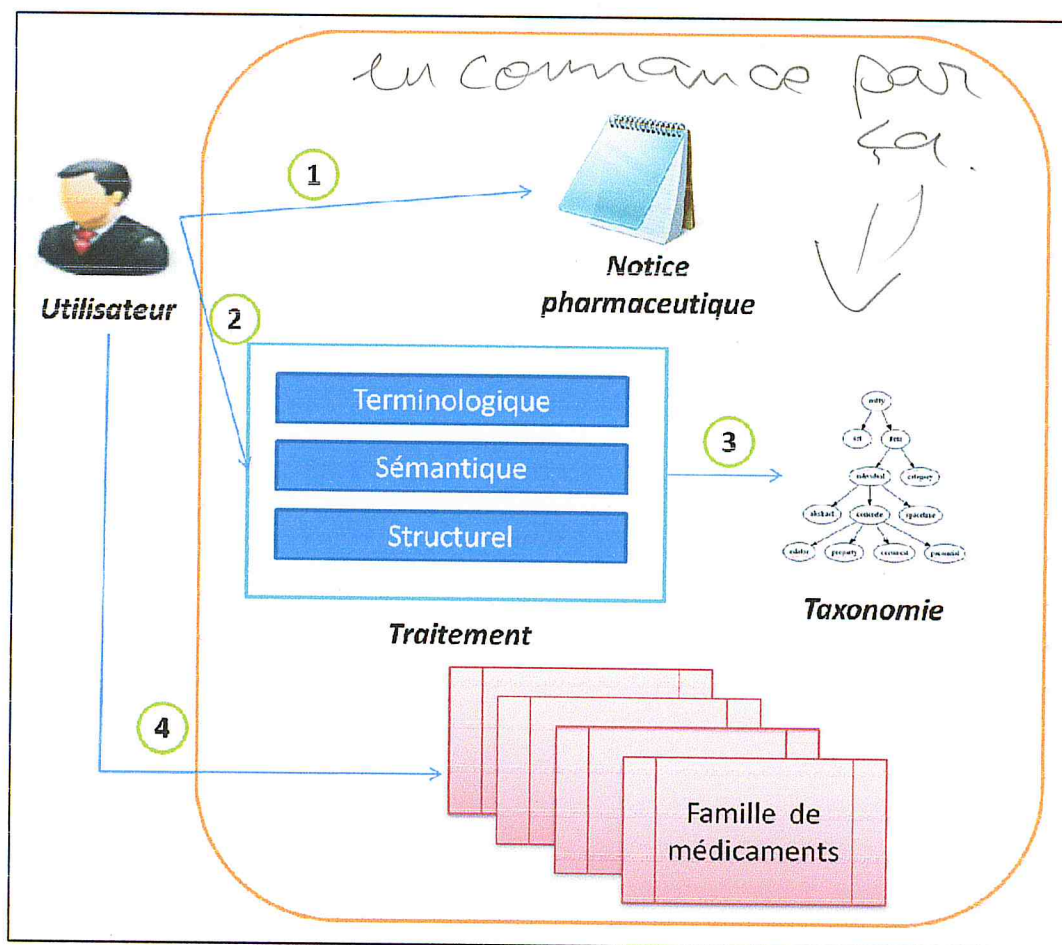


Figure 6: Architecture du système CSCNPharm

- 1- Le processus commence par l'utilisateur qui sélectionne une notice pharmaceutique,
- 2- Une fois la notice sélectionnée, l'utilisateur lance la batterie de traitement (terminologique, sémantique et structurel),
- 3- Après avoir vu les statistiques de traitement linguistiques (nombre de mots pertinents, mots clés, ...), l'utilisateur lance l'étape de cartographie qui va agir en transformant les vues en se basant sur les structures construites,
- 4- Si l'utilisateur est satisfait du résultat de la cartographie, alors il la sauvegarde (sous forme d'ontologie) qui va se classer automatiquement dans la catégorie qui convient grâce à l'étape de traitement sémantique faite préalablement.

5. Conclusion

Dans ce chapitre, nous avons décrit les motivations qui nous ont poussé à adopter certaines démarche dans l'optique de modélisation d'une approche personnalisée pour la cartographie sémantique et la classification de notices pharmaceutiques. Dans le dernier chapitre de ce mémoire, nous allons présenter quelques algorithmes implémentés et outils utilisés mais aussi justifier certains choix d'ordre pratiques.

Chapitre V :

Implémentation

du système

1. Introduction

Après avoir effectué la conception de CSCNPharm, nous allons dans cette section, présenter, dans un premier lieu, l'environnement de développement (langages et outils) ensuite, le diagramme d'accessibilité du système, quelques algorithmes, et enfin nous présenterons quelques captures d'écran.

2. Environnement de développement

2.1. Langage utilisé

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton de Sun Microsystems.

Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitation (Windows, Linux, Macintosh, Solaris).

Le langage Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut-être utilisé sur internet pour des petites applications intégrées aux pages web (applet) ou encore comme langage serveur (JSP).

2.2. OUTILS

2.2.1. Eclipse

Eclipse IDE est un environnement de développement intégré libre, le terme *Eclipse* désigne également le projet correspondant, lancé par IBM. Il est extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java à l'aide de la bibliothèque graphique SWT, d'IBM.

2.2.2. LES APIS

- **Jena:** est un ensemble d'outils (une API) permettant de lire et de manipuler des ontologies décrites en OWL et d'y appliquer certains mécanismes d'inférences. Au cours de notre développement, on a utilisé la version Jena2.6 qui permettait entre autres : la création et l'extraction de concepts, de propriétés (relations et attributs) sur les concepts ainsi que les restrictions sur les propriétés mais

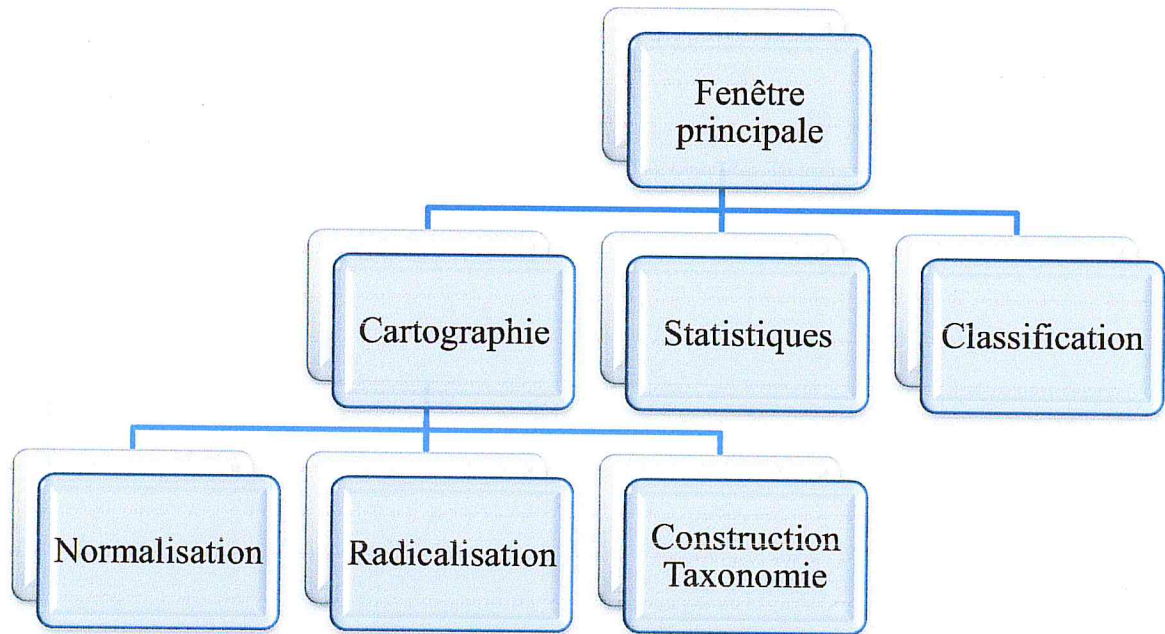


Figure 7: Diagramme d'accessibilité du système CSCNPharm.

4. Quelques algorithmes

Pour l'application de nos algorithmes, nous avons constitué un corpus de 20 notices pharmaceutiques (obtenues grâce au logiciel Vidai Expert) transformées au format texte (*.txt) et appartenant à quatre familles de médicaments différents, antalgiques, anti-inflammatoires, appareil-respiratoires, ophtalmiques. Ce choix se justifie par le fait que certaines de ces familles partagent des caractéristiques communes et certaines possèdent des caractéristiques diamétralement opposées les unes aux autres. Le but étant de tester la validité de l'algorithme de classification.

4.1. Cartographie

Algorithme général de cartographie :

Début

```
Ouvrir (notice) ; //la notice est au format *.txt  
Tableau de String TERMES = Extraction (notice) ;  
String famille = rechercheFamille(notice) ;  
Tableau de String MotsPertinents = Normalisation (TERMES);  
Ontologie taxonomie = ConstructionOntologie (MotsPertinents) ;
```

Fin ;

4.2. Extraction

Extraction (n: notice) : Tableau de String

Début

- Déclarer une liste des signes de ponctuation et des symboles inutile ;
- //les symbole inutile sont : ({} []|~@#\$\$%^&*()_ -+=<>^!~) ...etc
- Découper le texte de notice «n » en 3 sections (DCI, Indications et effets indésirables) ;
- Supprimer les accents, majuscules et les points dans les acronymes.
- Stocker cette liste dans un tableau intermédiaire nommé TERMES ;
- Trier ce tableau et classer les mots par ordre alphabétique ;
- Retourner (TERMES) ;

Fin:

4.3. Normalisation

Normalisation (TERMES : Tableau de String) : Tableau de String

Début

```
Déclarer une liste des mots vide nommé LV ;
Pour chaque mot de tableau TERMES
Faire
    Si (le mot est dans LV)
        | Alors éliminer le mot ;
        | Sinon garder et stocker le mot dans un tableau ;
    fsi;
Fait ;
Transformer les mots retenu en forme canonique ;
Début
    | Réduction du pluriel des noms à la forme au singulier ;
Fin ;
Stocker le résultat dans un tableau nommé MotsPertinents ;
Retourner (MotsPertinent) ;
```

Fin:

4.4. Recherche famille

RechercheFamille (n: notice) : String

Début

```
famille : String
//Déclarer quatre tableaux de String pour quatre familles de médicaments
//Chacun des tableaux contient les type médicaments de la famille
Ouvrir (notice) ;
Chercher la DCI dans la notice
Faire
    Comparer la DCI à tous les tableaux
    Si (DCI ∈ à un des tableaux)
        | Alors famille = nom de la famille;
        | Sinon tester la famille suivante ;
    fsi;
Fait ;
Retourner (famille) ;
```

Fin:

4.5. Construction Ontologie

constructionOntologie (MotsPertinents : Tableau de String) : Ontologie

Début

- Chercher pour chacune des trois sections de la notice les relations de dépendances entre les mots pertinents,
Affecter la DCI à la racine de l'Ontologie,
- Construire progressivement (selon les relations) les deux sous arbres « Indications » et « effets Indesirables » et les relier à la racine
- Retourner l'ontologie finale (au format *.owl) ;

Fin :

5. Quelques captures d'écran

La fenêtre principale du logiciel se présente comme suit (figure 8) :

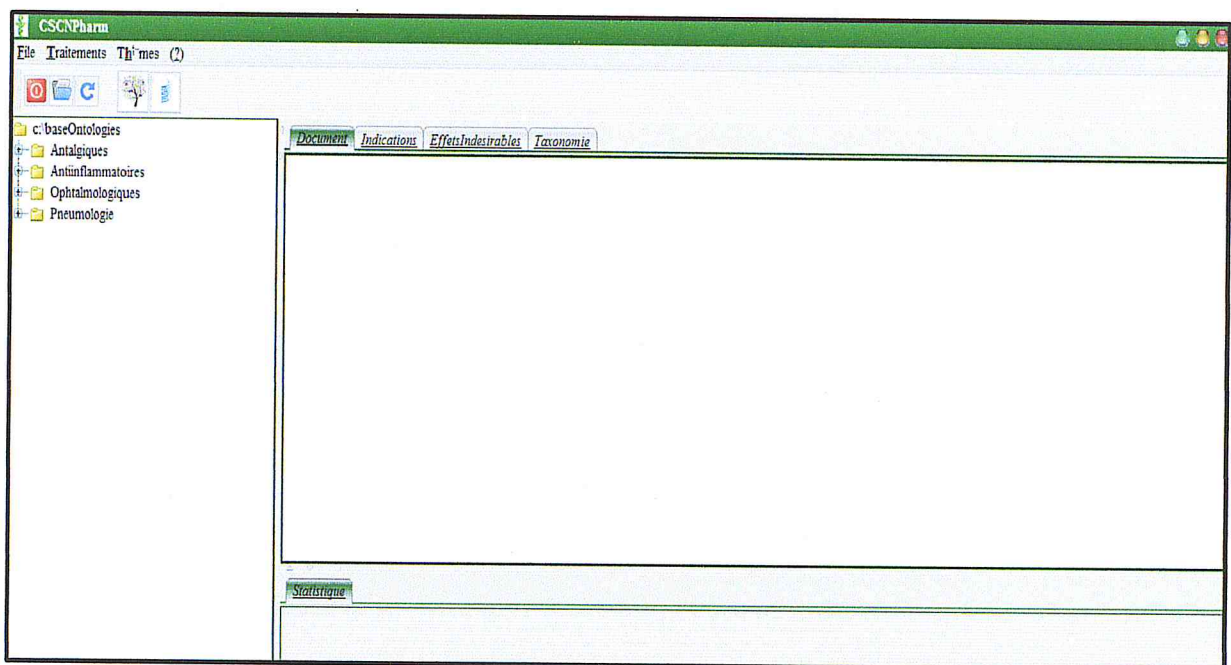


Figure 8: Fenêtre principale de CSCNPharm

Elle contient en haut, les différents menus et outils nous permettant d'effectuer les opérations de traitement terminologique, de cartographie et de classification. La partie centrale de la fenêtre, permet quant à elle de visualiser la notice, ainsi que sa segmentation en trois parties mais aussi la réduction de ces dernières après nettoyage (par extraction et normalisation) et la visualisation de la taxonomie finale. Par ailleurs, la partie gauche de la fenêtre présente une arborescence de nos quatre familles tests. Enfin le volet bas de la fenêtre, est réservé à l'affichage de statistiques liées aux opérations de traitement terminologique.

On sélectionne une notice de notre basesNotices (figure 9):

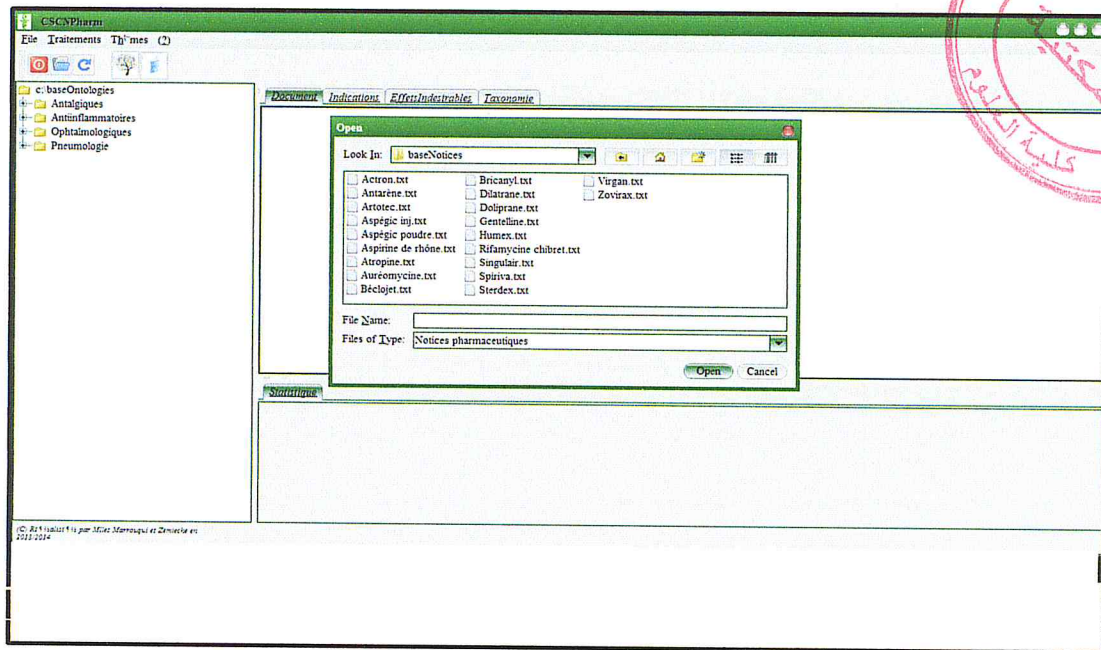
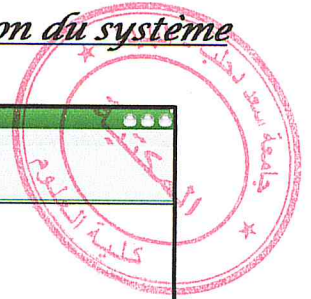


Figure 9: Fenêtre d'ouverture d'une notice

Quand on ouvre une notice (soit la notice Antarene), cette dernière s'affiche dans l'onglet « Document », comme illustré dans la figure 10 :

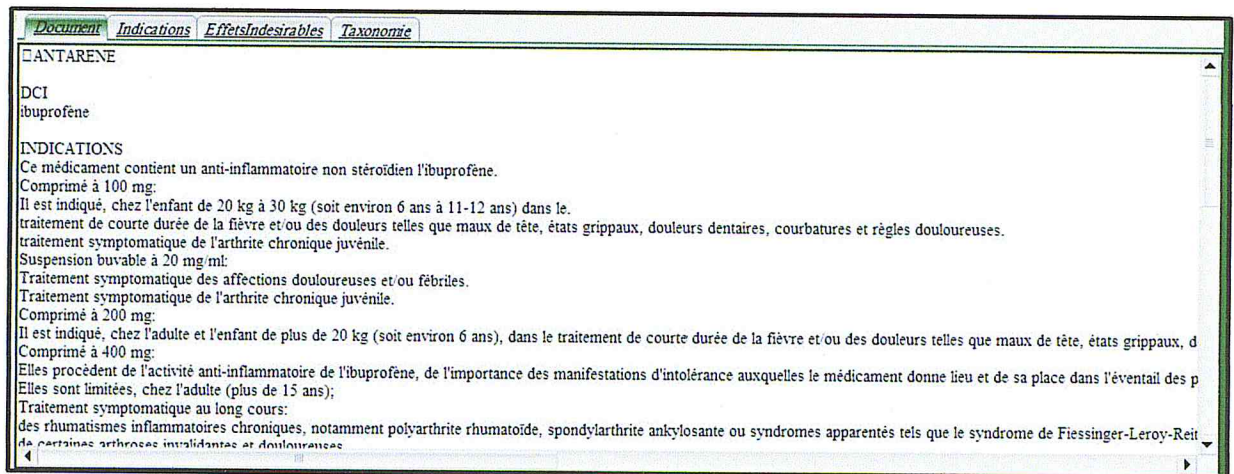


Figure 10: Affichage de la notice

Un clic sur le bouton *segmentation*, déclenche la subdivision du document en trois parties et l'affichage des parties *indications* et *effets indésirables* dans leurs onglets respectifs (figure 11).

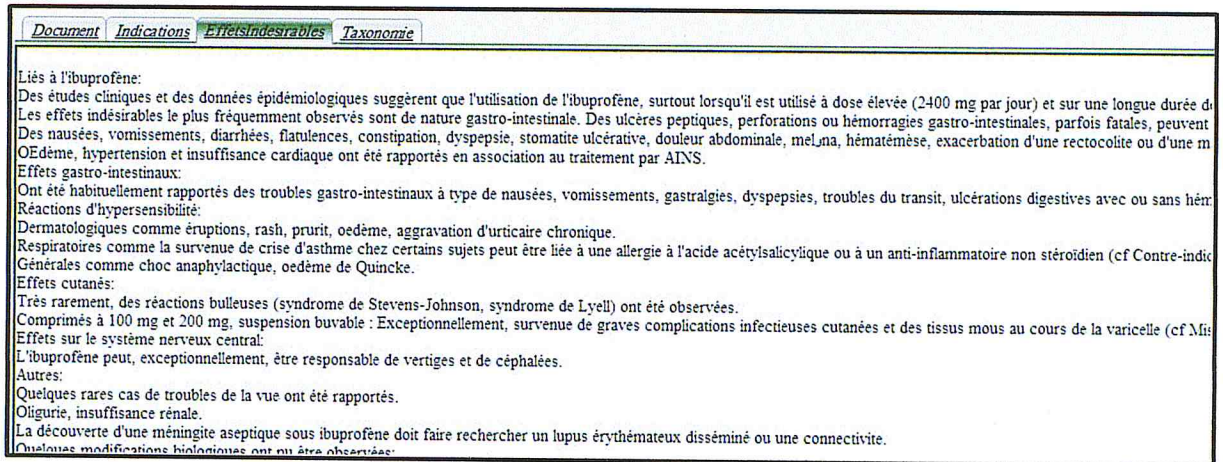


Figure 11: Etape de segmentation.

Quand on clique sur le bouton *extraction*, on remarque l'alignement de tous les mots de la notice choisie. Ainsi, tous les mots de la notice sont transformés en minuscule et purifiés des accents s'ils en contiennent (figure 12).

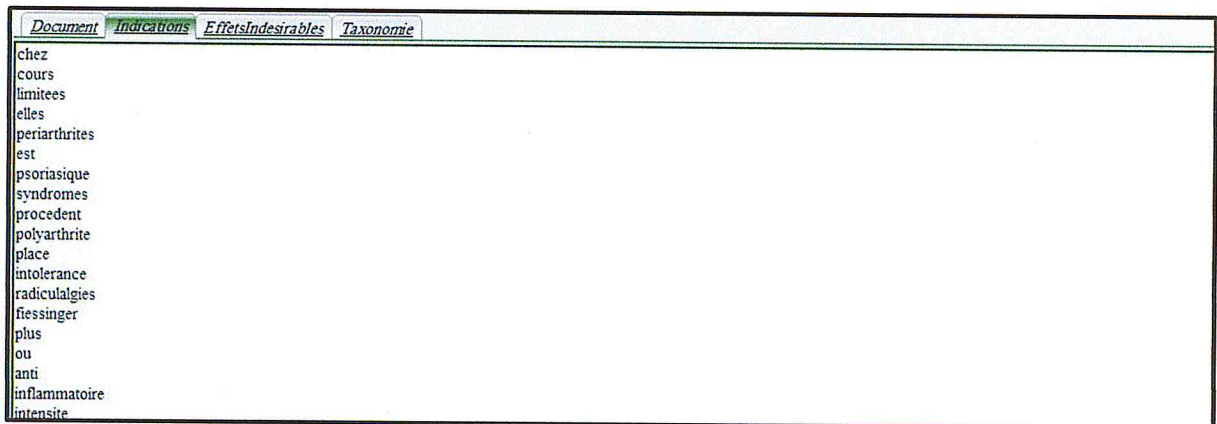


Figure 12: Etape d'extraction.

Quand on clique sur le bouton *normalisation*, on remarque l'élimination de tous les mots vides, ainsi que les doublons. Pour ne laisser au final que les mots pertinents (figure 13):

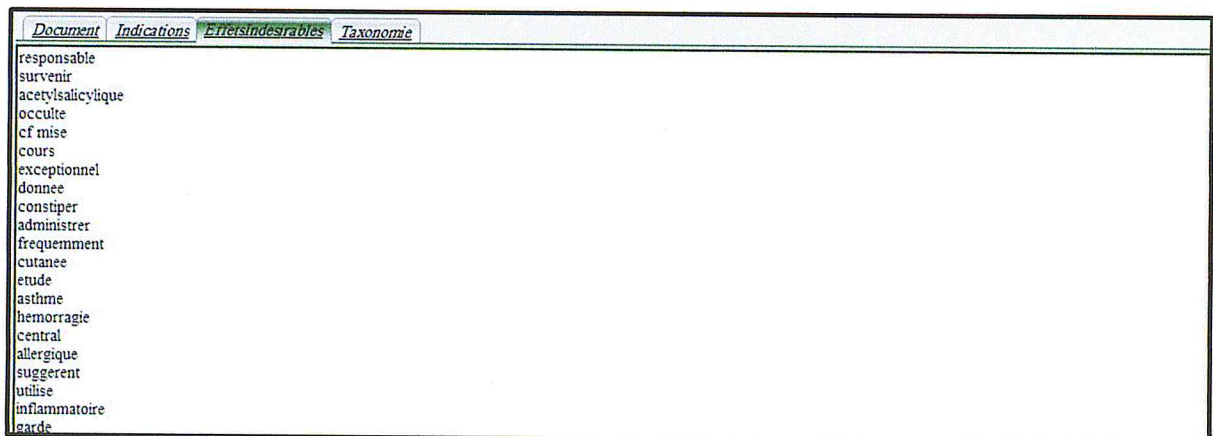


Figure 13: Etape de normalisation.

Par ailleurs, des informations d'ordre statistiques s'affichent dans l'onglet « statistiques », comme illustré dans la figure 14 :

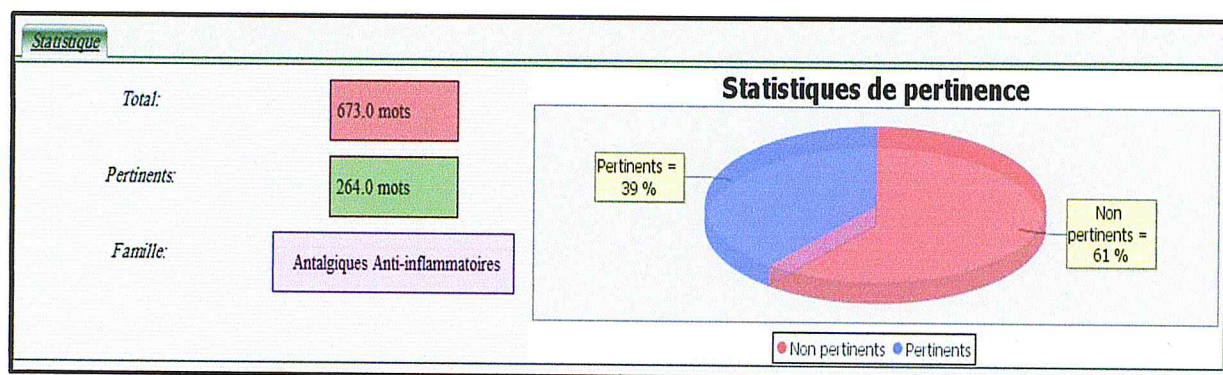


Figure 14: Fenêtre des statistiques.

Ainsi, nous pouvons remarquer qu'au départ, la notice contenait 673 mots qui après traitement terminologique (extraction, et normalisation) ont été réduit à 264 mots considérés comme étant pertinents (soit 39% de mots pertinent par rapport au contenu total de la notice). Nous pouvons par ailleurs, remarquer que comme pronostic avant classification, notre algorithme a détecté que le médicament appartient à deux familles de médicaments différents (Antalgiques et Anti-inflammatoires).

Un clic sur le bouton *cartographie*, permet d'afficher la taxonomie de concepts relative à la notice sélectionnée, comme illustré dans la figure 15 :

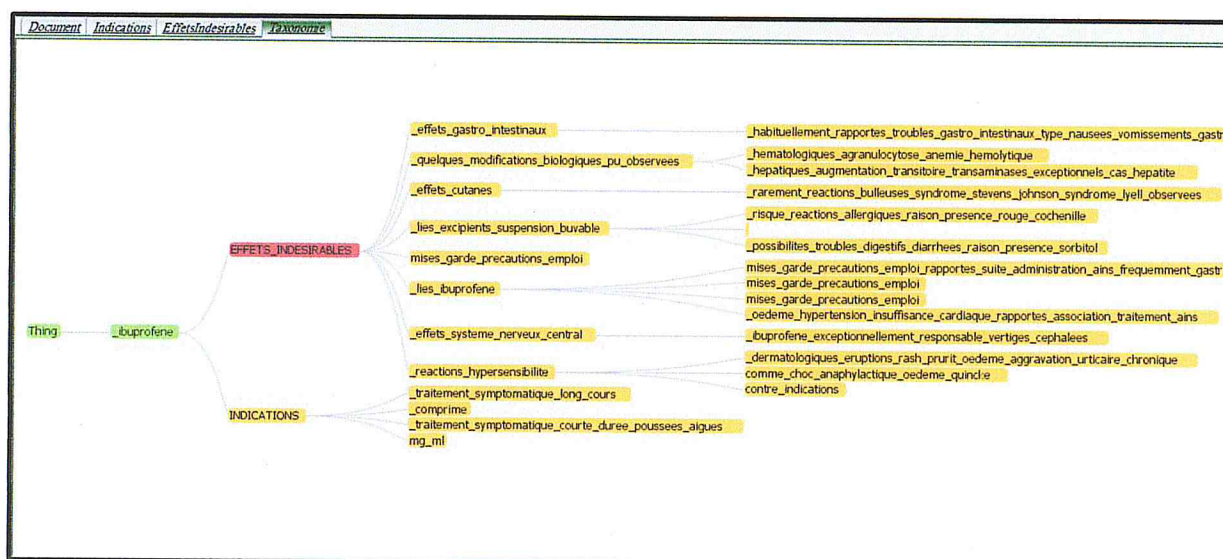


Figure 15: Visualisation graphique (taxonomie).

Enfin, si l'utilisateur est satisfait de la visualisation, il clique sur le bouton *classification*, et l'ontologie relative à la taxonomie se classera automatiquement dans la/ les famille(s) détectées, comme illustré dans la figure 16 :

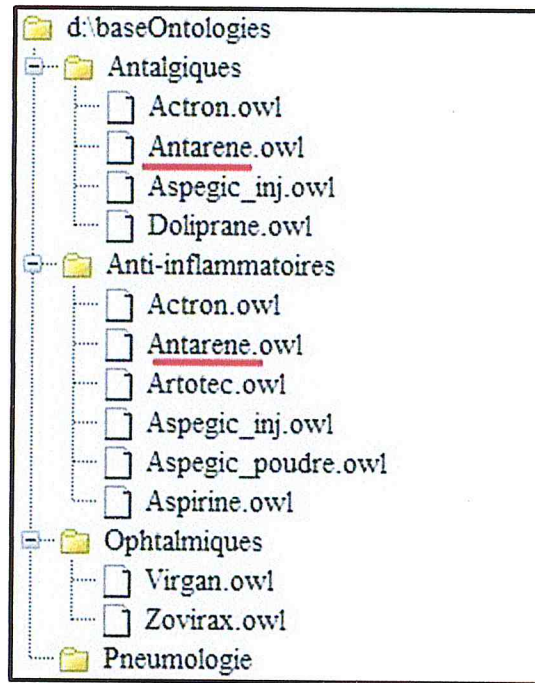


Figure 16: Classification d'une notice

5. Tests et validation du système

5.1. Corpus de tests

En l'absence de bases de notices -Algériennes- au format numérique et vu l'incapacité des outils de reconnaissances de caractères à extraire de manière optimale le contenu des notices pharmaceutique scannées, nous avons eu recours à un logiciel de gestion de pharmacie Français Vidal qui nous a permis de constituer un petit corpus de 20 notices pharmaceutiques. Le choix des notices était motivé par le choix de quatre familles que nous avons jugé particulièrement intéressante pour tester nos algorithmes de cartographie et de classification.

Les familles que nous avons choisi sont : Les *Antalgiques* et les *Anti-inflammatoires*, qui partagent quelques similitudes et donc la capacité de notre algorithme de classification à distinguer entre des médicaments appartenant à une de ces deux familles ou à l'autre serait un bon indicateur de performance. Par ailleurs, nous avons choisit deux autres familles de médicaments, à savoir : les médicaments destinés aux maladies *Ophtalmiques* et *Pneumologiques*. Telles qu'elles représentent des caractéristiques disjointes et donc le succès de l'algorithme à filtrer les médicaments appartenant à ces deux classes prouverait sa performance en termes d'appréhension de classes différentes et donc sa capacité à être généralisé pour d'autres champs d'études que le domaine pharmaceutiques. Enfin, les différentes notices (dont le contenu est indéniablement différents, différentes indications, différents effets indésirables et différentes structurations et présentations) nous ont permis de tester notre algorithme de cartographie et sa capacité à appréhender des informations de différentes granularités sémantiques.

5.2. Mesures de performance utilisées

Pour nous assurer de l'optimalité des résultats de classification fournis par CSCNPharm, nous avons soumis le résultat de classification à une évaluation et ce en ayant recours à quelques mesures comme : précision, rappel, overall et f-mesure (dites mesures de performance du système).

1. Définition des mesures de performance utilisées

Les mesures utilisées pour évaluer la qualité de la classification obtenue sont principalement les mesures de calcul de la pertinence en recherche d'information, telles que la *précision* et le *rappel*.

Le calcul de ces mesures [20], est basé sur la comparaison entre les classifications produites par un système automatique qu'on appellera S et un ensemble de classification de référence produit par un humain qu'on notera H .

- Les classifications *correctes trouvées* par un système sont appelées « the true positives (TP) » et sont calculées ainsi :

$$TP = S \cap H$$

- Les classifications *incorrectes trouvées* par un système sont appelées « the false positives (FP) » et sont calculées ainsi :

$$FP = S - S \cap H$$

- Les classifications *correctes omises* par un système sont appelées « the false négatives (FN) » et sont calculées ainsi :

$$FN = H - S \cap H$$

- La *précision* est une mesure d'exactitude, elle varie entre [0,1] elle est calculée de la manière suivante :

$$Precision = \frac{|TP|}{|TP + FP|} = \frac{S \cap H}{S}$$

- Le *rappel* est une mesure de perfection, elle varie entre [0,1], elle est calculée de la manière suivante :

$$Rappel = \frac{|TP|}{|TP + FN|} = \frac{S \cap H}{H}$$

En fait, le *rappel* peut prendre des valeurs importantes aux dépens de la *précision*, en retournant toutes les classifications possibles. En même temps, la *précision* peut prendre des valeurs importantes aux dépens du *rappel*, en retournant que les classifications correctes cependant peu nombreuses.

- C'est pour ces raisons qu'il est préférable de prendre en considération les deux mesures simultanément via une mesure qui combine le rappel et la précision telles que : la *F-mesure* qui se calcule de la manière suivante :

$$F - \text{Mesure} = \frac{2 * (\text{Rappel} * \text{Précision})}{\text{Rappel} + \text{Précision}}$$

La *F-mesure* est une mesure globale de la qualité des classifications produites, elle varie entre [0,1], cette mesure alloue la même importance à la précision et au rappel.

- Une autre mesure de la qualité de classification et qui combine le rappel et la précision : *l'overall* qui se calcule de la manière suivante :

$$\text{OVERALL} = \text{Rappel} \left(2 - \left(\frac{1}{\text{Précision}} \right) \right)$$

L'overall peut prendre des valeurs négatives si le nombre de fausses classification (FP) trouvées par le système dépasse le nombre de classifications correctes (TP) trouvées par le système.

Dans la plupart des cas *l'overall* est plus petit que le rappel et la précision, ce qui rend difficile d'atteindre un *overall* supérieur à 0.5.

- Une mesure pour évaluer le pourcentage d'erreurs du système automatique est le *Fallout* qui se calcule de la manière suivante :

$$\text{Fallout} = \frac{FP}{FP + TP}$$

5.3. Jeux de données

Afin de tester la validité de notre système nous avons adopté le même procédé, expliqué ci-dessus avec la notice du médicament *Antarène*.

Le tableau suivant (tableau 3) montre la classification manuelle et celle de CSCNPharm des vingt notices sélectionnées.

Notices	Classification manuelle	Classification CSCNPharm
Actron	Antalgiques Anti-inflammatoires	Antalgiques Anti-inflammatoires
Antarène	Antalgiques Anti-inflammatoires	Antalgiques Anti-inflammatoires
Artotec	Anti-inflammatoires	Anti-inflammatoires
Aspegic injection	Antalgiques Anti-inflammatoires	Antalgiques Anti-inflammatoires
Aspegic poudre	Anti-inflammatoires	Anti-inflammatoires

Aspirine	Anti-inflammatoires	Anti-inflammatoires
Atropine	Ophtalmiques	Ophtalmiques
Aureomycine	Ophtalmiques	Ophtalmiques
Beclojet	Pneumologie	Pneumologie
Bricanyl	Pneumologie	Pneumologie
Dilatrane	Pneumologie	Pneumologie
Doliprane	Antalgiques	Antalgiques
Gentelline	Ophtalmiques	✘
Humex	Pneumologie	Pneumologie
Rifamycine chibret	Ophtalmiques	Ophtalmiques
Singulair	Pneumologie	Pneumologie
Spiriva	Pneumologie	Pneumologie
Sterdex	Ophtalmiques	Ophtalmiques
Virgan	Ophtalmiques	Ophtalmiques
Zovirax	Ophtalmiques	Ophtalmiques
<i>Classification du Système : S = 19</i> <i>Classification de l'humain : H = 20</i>		

Tableau 3. Classification manuelle / CSCNPharm (test et validation).

Dans le tableau 3, nous remarquons que le système a échoué à classer une seule notice *Gentelline* et de ce fait les mesures de performances du système en terme de classification se présentent comme suit :

TP	FP	FN	Précision	Rappel	F-mesure	OVERALL	Fallout
19	0	1	1	0,95	0,97	0,95	0

Tableau 4. Mesures de performances de CSCNPharm

Comme le montre le tableau 4, les résultats de l'algorithme de classification sont plutôt satisfaisants. Par ailleurs, en ce qui concerne les résultats de cartographie, nous nous sommes basés sur notre expérience dans le domaine pharmaceutique et constatons que les visualisations offertes par le système, offrent dans la plupart des cas, le degré d'information nécessaire à une activité décisionnelle aisée dans le domaine pharmaceutique.

Conclusion et perspectives

Conclusion et perspectives

- Souplesse dans l'activité décisionnelle. En effet, les informations fournis par le système permettront à l'utilisateur de décider sur la faisabilité de l'administration du médicament ou pas à un patient donné ;
- Ergonomie : Son interface très malléable, rend CSCNPharm très agréable d'utilisation. Par ailleurs, l'utilisation des diagrammes statistiques et de visualisations colorées apporte un apport considérable à l'activité décisionnelle ;
- Extensibilité, Son caractère extensible lui confère un côté relativement générique de par sa faible dépendance vis-à-vis du domaine d'application. En effet, bien qu'appliquer dans le cadre de gestion de pharmacie, CSCNPharm aurait très bien pu être utilisé dans les laboratoires ou en écoles de pharmacie pour que les chercheurs et étudiants novices s'habituent aux différentes propriétés des médicaments et les dangers quant à leur éventuelle interaction.

2. Perspectives

Durant les quelques mois consacrés à la réalisation de notre projet de fin d'études nous nous sommes obstinées à atteindre les objectifs qui nous étaient fixés au départ en y ajoutant d'autres idées. Toutefois, dans un souci d'amélioration du système proposé et d'étoffer le spectre de ses fonctionnalités, nous proposons les perspectives suivantes :

- Étendre la cartographie aux différentes parties de la notice afin d'offrir une vision globale et enrichir l'activité décisionnelle;
- Envisager un corpus de notices plus grand avec un nombre de familles plus important afin de pouvoir raffiner les tests de classifications ;
- Tirer profit des Ontologies de domaine (médical ou pharmaceutique) pour enrichir le contenu des notices ce qui aura pour effet de renforcer la dimension sémantique ;
- Avoir recours à des ressources dictionnaires externe comme WordNet qui permettront d'approfondir le traitement lexical afin de découvrir ou corriger les relations entre les différents mots pertinents.

1. Conclusion

L'ère de l'information que nous connaissons aujourd'hui a été marquée par deux phénomènes importants. Le premier, d'ordre technologique, concerne la généralisation des documents numériques. Le second, d'ordre économique, concerne la reconnaissance du fait que la compétitivité économique réside dans la maîtrise des flux d'informations. Ainsi, l'accès à l'information est devenu un enjeu stratégique mais encore faut-il bien comprendre et appréhender l'information contenue dans les documents numériques.

On peut dire que la cartographie sémantique se présente comme un Système de Gestion des Connaissances qui peut être vu comme une identification du patrimoine de connaissances qui permet aux organisations désireuses de gérer leur patrimoine de connaissances, d'en faire une analyse fine afin de déterminer, dans leur stratégie, quelles sont les connaissances qu'elles doivent pérenniser, développer, abandonner, etc. La cartographie devient alors un outil d'aide à la décision.

Les travaux menés dans ce mémoire nous ont permis d'approfondir nos connaissances dans le domaine de cartographie qui lui-même appartient au domaine de l'ingénierie des connaissances. Notre objectif a été de tirer profit des travaux menés dans cette voie et nous nous sommes intéressées plus particulièrement aux méthodes de visualisation hiérarchique des données tout en modifiant le degré de pertinence et de réutilisation de la cartographie générée. Nos recherches et constatations, nous ont conduit à la mise au point d'une méthode originale nommée CSCNPharm (Cartographie Sémantique et Classification des Notices Pharmaceutiques). L'originalité de notre méthode réside dans le domaine d'application en question car considéré comme étant *critique* mais aussi dans la combinaison des caractéristiques suivantes :

- Automatisation totale du processus de cartographie et de classification, l'utilisateur n'aura qu'à sélectionner la notice qu'il veut visualiser et classer et le système, le décharge entièrement des tâches superflues comme l'intervention pour la rectification de la visualisation ;
- Intégration: le système s'intègre facilement dans n'importe quel environnement car il porte en son sein toutes les API essentielles, et ressources externes nécessaires à son bon fonctionnement ;

6. Conclusion

Dans ce chapitre, nous avons présenté les outils utilisés, les algorithmes mis au point, ainsi que quelques captures d'écran de notre système de cartographie et classification des notices pharmaceutiques.

Annexe A :
Généralités sur les
Ontologies

Annexe A : généralités sur les Ontologies

3. Origine des ontologies

Le terme *Ontologie* a été utilisé pour la première fois par les philosophes grecs dans une discipline qui a plus 2300 ans, qui s'intéresse à l'existence de l'être en tant qu'être et aux catégories fondamentales de l'existant. Etymologiquement parlant, *Onto* est un terme dérivé d'un mot grec et signifie l'Être (ainsi que ses manifestations) *Logie* vient du mot grec *Logos* qui veut dire Science ou Étude. Donc, l'Ontologie est la science qui s'intéresse à l'étude de l'être en tant qu'être.

Dans la philosophie classique, l'Ontologie correspond à ce qu'Aristote appelait la Philosophie première, c'est-à-dire *la science de l'être en tant qu'être*, par opposition aux philosophies secondes qui s'intéressent à l'étude des manifestations de l'être (les *existants*).

Ensuite la notion d'ontologie a été abordée dans le domaine de l'intelligence artificielle (IA) par John McCarthy. Il affirmait déjà en 1980 que les concepteurs des systèmes intelligents fondés sur la logique devraient d'abord énumérer tout ce qui existe.

1. Qu'est-ce qu'une ontologie ?

1.1. Définitions

Il est difficile de définir ce qu'est une ontologie d'une façon définitive. Le mot est en effet employé dans des contextes très différents touchant la philosophie, la linguistique ou l'intelligence artificielle. Nous allons nous intéresser à sa définition dans le domaine informatique.

Dans le cadre de l'intelligence artificielle, **Neeches** et ses collègues ont été les premiers à proposer une définition à savoir : «*Une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire*».

En 1993, **Gruher** propose la définition suivante : «*spécification explicite d'une conceptualisation*» qui est jusqu'à présent la définition la plus citée dans la littérature en intelligence artificielle. Cette définition a été modifiée légèrement par **Borst** comme «*spécification formelle d'une conceptualisation partagée*». Ces deux définitions sont regroupées dans celle de **Studer** comme «*Spécification formelle et explicite d'une conceptualisation partagée*».

Annexe A : généralités sur les Ontologies

- *Formelle* : l'ontologie doit être lisible par une machine, ce qui exclut le langage naturel.
- *Explicite* : la définition explicite des concepts utilisés et des contraintes de leur utilisation.
- *Conceptualisation* : le modèle abstrait d'un phénomène du monde réel par identification des concepts clefs de ce phénomène.
- *Partagée* : l'ontologie n'est pas la propriété d'un individu, mais elle représente un consensus accepté par une communauté d'utilisateurs.

Pour **Guarino & Giaretta** «*Une ontologie est une spécification rendant partiellement compte d'une conceptualisation*».

Swartout et ses collègues la définissent comme suit : «*une ontologie est un ensemble de termes structurés de façon hiérarchique, conçue afin de décrire un domaine et qui peut servir de charpente à une base de connaissances*».

La même notion est également développée par **Gomez** comme suit : «*une ontologie fournit les moyens de décrire de façon explicite la conceptualisation des connaissances représentées dans une base de connaissances*».

Pour conclure, nous pouvons donc constater que les définitions, dans leur diversité, offrent des points de vue à la fois différents et complémentaires sur un même concept. En clair, une ontologie est un schéma particulier qui décrit, à l'aide d'un vocabulaire structuré et un langage formel non ambigu, les concepts et les propriétés pertinents pour un domaine d'application donné.

1.2. Termes relatifs

- **Otologiste** : Personne qui exploite les principes d'ontologies, soit pour les construire ou parce que son travail est en relation avec elles.
- **Thésaurus** : Un ensemble de mots clés (descripteurs) hiérarchisés relatifs à un domaine donné et les relations entre eux (is-a, synonymes, antonymes, ...etc)
- **Les taxonomies**: Une taxonomie est la forme la plus simple (représentation graphique) des ontologies. Elle se présente par une hiérarchie des termes généralement organisés dans le sens de la spécialisation (du général au particulier) bien que d'autres organisations

Annexe A : généralités sur les Ontologies

soient possibles mais on ne peut représenter qu'une seule relation à la fois.

2. Composants d'une ontologie

Comme nous l'avons dit, les ontologies fournissent un vocabulaire commun d'un domaine et définissent la signification des termes et des relations entre elles. La connaissance dans les ontologies est principalement formalisée à l'aide de cinq éléments: *Concepts, Relations, Fonctions, Axiomes et Instances*.

2.1. Les concepts

Un concept est un principe, une idée, une notion abstraite, sémantiquement évaluable et communicable. L'ensemble des propriétés d'un concept constitue sa compréhension ou son intension, et l'ensemble des termes qu'il englobe et son extension.

Un concept est composé de trois parties :

- *Un (ou plusieurs) terme(s)* : c'est la représentation linguistique des concepts, les termes permettent de désigner le concept. Ces termes sont aussi appelés *labels* de concepts.

Synonymic : plusieurs termes dénotent le même concept

– Ambiguïté : plusieurs concepts dénotés par le même terme □

- *Une notion* : elle correspond à la sémantique du concept, elle est définie à travers ses propriétés et ses attributs. Elle est appelée *intention* du concept.

- *Un ensemble d'objets* : il correspond aux objets définis par le concept, il est appelé *extension* du concept. Les objets sont les *instances* du concept.

2.1.1. Les propriétés portant sur un concept

- *La généralité* : un concept est générique s'il n'admet pas d'extension

Exemple : la vérité est un concept générique.

- *L'identité* : un concept porte une propriété d'identité si cette propriété permet de conclure quant à l'identité de deux instances de ce concept. Cette propriété peut porter sur des attributs du concept ou sur d'autres concepts.

Exemple : le concept d'étudiant porte une propriété d'identité liée au numéro de l'étudiant, deux étudiants étant identiques s'ils ont le même numéro.

Annexe A : généralités sur les Ontologies

- *La rigidité* : un concept est dit rigide si toute instance de ce concept en reste instance dans tous les mondes possibles.

Exemple : humain est un concept rigide, étudiant est un concept non rigide.

- *L'anti-rigidité* : un concept est anti-rigide si toute instance de ce concept est essentiellement définie par son appartenance à l'extension d'un autre concept.

Autrement dit : un concept est *anti rigide* s'il peut être une instance pour d'autres concepts.

Exemple : étudiant est un concept anti-rigide car l'étudiant est avant tout un humain.

2.1.2. Les propriétés portant sur deux concepts

- *L'équivalence* : deux concepts sont équivalents s'ils ont la même extension.

Exemple : étoile du matin et étoile du soir.

- *La disjonction* : (on parle aussi d'incompatibilité) deux concepts sont disjoints si leurs extensions sont disjointes.

Exemple: homme et femme.

- *La dépendance* : Un concept C1 est dépendant d'un concept C2 si pour toute instance de C1 il existe une instance de C2 qui ne soit ni partie ni constituant de l'instance de C1.

Exemple : parent est un concept dépendant de enfant (et vice-versa).

2.2. Les relations

Elles représentent des interactions entre concepts permettant de construire des représentations complexes de la connaissance du domaine. Elles établissent des liens sémantiques binaires, organisables hiérarchiquement.

2.2.1. Les propriétés fondamentales à une relation

- *Les propriétés algébriques* : symétrie, réflexivité, transitivité
- *La cardinalité* : nombre possible de relations de ce type entre les mêmes concepts (ou instances de concept). Les relations portant une cardinalité représentent souvent des attributs.

Exemple: une pièce a au moins une porte.

2.2.2. Les propriétés liant deux relations

Annexe A : généralités sur les Ontologies

- *L'incompatibilité* : on dit que deux relations R1 et R2 sont incompatibles si elles ne peuvent lier les mêmes instances de concepts.

Exemple : « être rouge » et « être vert » sont deux relations incompatibles.

- *L'inverse* : on dit que deux relations binaires R1 et R2 sont inverses l'une de l'autre si, quand R1 lie deux instances I1 et I2, alors R2 lie I2 et I1.

Exemple : « a pour père » et « a pour enfant » sont deux relations inverses l'une de l'autre.

- *L'exclusivité* : deux relations R1 et R2 sont exclusives si, quand R1 lie des instances de concepts, R2 ne lie pas ces instances, et vice-versa. L'exclusivité entraîne l'incompatibilité.

Exemple : l'appartenance et la non appartenance sont exclusives.

2.3. Les fonctions

Elles constituent des cas particuliers de relation, dans laquelle le nième élément de la relation, est défini en fonction des n-1 éléments précédents.

2.4. Les axiomes

Les axiomes sont des expressions qui sont toujours vraies. Ils ont pour but de définir dans un langage logique la description des concepts et des relations. Leur inclusion dans une ontologie peut avoir plusieurs objectifs:

- Définir la signification des composants.
- Définir des restrictions sur la valeur des attributs.
- Définir les arguments d'une relation.
- Vérifier la validité des informations spécifiées ou en déduire de nouvelles.

2.5. Les instances (individus)

Elles constituent la définition extensionnelle de l'ontologie ; elles sont utilisées pour représenter des éléments dans un domaine.

Exemple: les individus *Imane* et *Manel* sont des instances du concept *Personne*.

3. Les langages d'ontologies

Il existe plusieurs langages de représentation des ontologies, les plus connus et les plus utilisés sont :

Annexe A : généralités sur les Ontologies

3.1. KIF

KIF est un langage basé sur les prédicats du premier ordre avec des extensions pour représenter des définitions et des méta-connaissances, la logique du premier ordre étant un langage de bas niveau pour l'expression d'ontologies. Une extension du langage KIF, ONTOLINGUA, est utilisée dans le serveur d'édition d'ontologies, ONTOLINGUA du même nom.

3.2. RDF/RDF Schéma

Le W3C a adopté le langage RDF (Ressource Description Framework) comme un des formalismes standards de représentation de connaissances sur le Web. Utilisant la syntaxe XML (Extended Markup Language) qui constitue déjà un standard, le RDF permet de décrire des ressources Web en termes de ressources, propriétés et valeurs. Une ressource peut être une page Web (identifiée par son URI, United Resource Identifier) ou une partie de page (identifiée par une balise). Les propriétés couvrent les notions d'attributs, relations ou aspects et servent à décrire une caractéristique d'une ressource en précisant sa valeur. Les valeurs peuvent être des ressources ou des littéraux. RDF dispose d'une sémantique formelle analogue à celle des graphes conceptuels, c'est-à-dire identique à celle d'un fragment de la logique du premier ordre. Un schéma de base incluant les primitives sémantiques généralement utilisées, a ainsi été ajouté au RDF et constitue ce qu'on appelle le RDF SCHEMA.

3.3. DAML + OIL

Dans l'optique d'une utilisation d'ontologies sur le Web, le langage RDF-S a été enrichi par l'apport du langage OIL (Ontology Interchange Language) qui permet d'exprimer une sémantique à travers le modèle des frames tout en utilisant la syntaxe de RDF-S. OIL offre de nouvelles primitives permettant de définir des classes à l'aide de mécanismes ensemblistes issus des logiques de description (intersection de classes, union de classes, complémentaire d'une classe). Il permet également d'affiner les propriétés de RDF-S en contraignant la cardinalité ou en restreignant la portée. Le langage OIL a été fusionné avec le langage DAML pour former le DAML+OIL. DAML (Darpa Agent Markup Language) est conçu pour permettre l'expression d'ontologies dans une extension du langage RDF. Il offre les

Annexe A : généralités sur les Ontologies

primitives usuelles d'une représentation à base de frames et utilise la syntaxe RDF. L'intégration de OIL rend possible les inférences compatibles avec les logiques de description, essentiellement les calculs de liens de subsumption.

3.4. OWL

La combinaison de RDF/RDF-S et de DAML+OIL a permis l'émergence d'OWL (Web Ontology Language), un langage standard de représentation de connaissances pour le Web. Développé par le groupe de travail sur le Web Sémantique du W3C, OWL peut être utilisé pour représenter explicitement les sens des termes des vocabulaires et les relations entre ces termes. OWL vise également à rendre les ressources sur le Web aisément accessibles aux processus automatisés, d'une part en les structurant d'une façon compréhensible et standardisée, et d'autre part en leur ajoutant des méta-informations. Pour cela, OWL a des moyens plus puissants pour exprimer la signification et la sémantique que XML, RDF, et RDF-S. De plus, OWL tient compte de l'aspect diffus des sources de connaissances et permet à l'information d'être recueillie à partir de sources distribuées, notamment en permettant la mise en relation des ontologies et l'importation des informations provenant explicitement d'autres ontologies.

Bibliographie

- [1] C. J. c. D. Martinc, «Recherche et analyse de l'information textuelle (tendances des outils linguistiques),» *Documentaliste Sciences de l'Information*, vol. 40, n° 11, pp. 14-24, 2003.
- [2] A. C. Cario, «Indexation et recherche conceptuelle de documents pédagogiques guidée par la structure de wikipedia,» France, 2011.
- [3] D. S. Lyne, «Relations sémantiques pour l'indexation automatique (Définition d'objectifs pour la détection automatique),» *Document numérique*, vol. 8, n° 13, pp. 135-155, 2004.
- [4] K. Laurent, «Accès sémantique aux bases de données documentaires, techniques symboliques de traitement automatique du langage pour l'indexation thématique et l'extraction d'information temporelle,» France, 2011.
- [5] C. Stéphane, «Technologies linguistiques et modes de représentation de l'information textuelle,» *Documentaliste Sciences de l'Information*, vol. 44, n° 11, pp. 30-39, 2007.
- [6] D. Aurélie, «Cartographie des connaissances et gestion des ressources humaines: exemple de l'ambiguïté cognitive des Systèmes de Gestion des connaissances,» *Système d'Information et Management*, vol. 12, n° 13, pp. 31-56, 2007.
- [7] LUXINNOVATION G.I.E, «Gestion des connaissances, savoir implicite et explicite,» L'Agence Nationale pour la Promotion de l'Innovation et de la Recherche, Luxembourg, 2008.
- [8] M. Boukraa, Écrivain, *Knowledge management*. [Performance]. 2012.
- [9] C. R. e. R. T. Jean Charlet, «Ingénierie des connaissances,» Equipe OAK, Laboratoire INRIA, Université Paris sud, Paris, 2001.
- [10] G. -. L. Jean-Christophe, «Outils de knowledge Management (Résumé de lecture),» Ecole centrale de Marseille France, Avril 2008.
- [11] P. d. e. d. m. Visualisation de l'information, «Solveig Vidal,» Centre National de la Recherche Scientifique, France, Mai 2006.
- [12] C. Tricot, «Cartographie sémantique, des connaissances à la carte,» Équipe Condillac "ingénierie des connaissances", Laboratoire d'informatique, Systèmes, Traitement de

l'Information et de la Connaissance, France, 2006.

- [13] I. Quentin, «Cartographier les connaissances,» 07 03 2012. [En ligne]. Available: <http://isabellequentin.wordpress.com/2012/03/07/cartographier-les-connaissances-lapproche-par-les-processus/>. [Accès le Mai 2013].
- [14] Leem, les entreprises du médicaments, «La notice, un agent de sécurité,» 06 05 2011. [En ligne]. Available: <http://www.leem.org/dossier/notice-un-agent-de-securite>. [Accès le Mai 2013].
- [15] Institut de Recherche et de Documentation en Economie de la Santé, «La politique du médicament en France,» www.irdes.fr, France, Juin 2013.
- [16] Pharmaciens Sans Frontières Comité International, «NOTIONS DE BASE SUR LES MÉDICAMENTS,» Pharmaciens Sans Frontières, 2004.
- [17] ILPGA Université Paris 3, «Secteur TAL Informatique,» 20 03 2003. [En ligne]. Available: <http://www.tal.univ-paris3.fr/cours/parcours/introtal.htm>. [Accès le Mai 2013].
- [18] R. M. D. Scala, *Arbre, TAD d'arbre binaire, parcours*, France: RM di scala, 2005.
- [19] P. Saint-Dizier, «Taxonomie - sémanticlopédie,» Juin 2013. [En ligne]. Available: <http://www.semantique-gdr.net/dico/index.php?title=Taxonomie&printable=yes>. [Accès le Juin 2013].
- [20] Martin RAJMAN, Romaric BESANÇON, Jean-Cédric CHAPPELIER, «Le modèle DISIR: Une approche à base de sémantique distributionnelle pour la recherche documentaire,» *ATALA*, vol. 41, n° 12, pp. 1-27, 2000.
- [21] Nawel NASSR, «Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes,» France, 09 décembre 2002.
- [22] Fatiha BOUBEKEUR-AMIROUCHE, «Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets,» France, 01 Juillet 2008.
- [23] Mustapha BAZIZ, Nathalie AUSSENAC-GILLES, Mohand BOUGHANEM, «Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche D'information,» *INFORSID*, n° 12-906855-19-7, pp. 121-134, 2003.
- [24] Yaël CHAMPCLAUX, «Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information,» France, 4 décembre 2009.
- [25] W. Nesrine ZEMIRLI, «Vers le développement d'un système de recherche d'information personnalisé intégrant le profil utilisateur,» France, 2004.

- [26] Vincent CLAVEAU, Romain TAVERNARD, Laurent AMSALEG, «Vectorisation des processus d'appariement document-requête,» chez *CORiA 2010*, Sousse - Tunisie, 2010.
- [27] Aurélien Max, *Indexation et Recherche d'Information -Introduction*, Université Paris-Sud 11, Orsay, 2009-2010.
- [28] Rami HARRATHI, «Recherche d'information conceptuelle dans les documents semi-structurés,» France, 29 Septembre 2010.

