

Ministère de l'enseignement supérieur et de la recherche scientifique

Université M'Hamed Bougarra Boumerdes

Faculté des sciences

Département informatique

**Mémoire de fin d'études**

Pour l'obtention du diplôme d'ingénieur d'état en informatique

**Option** : système d'information avancé

**Thème :**

**Indexation automatique des documents électroniques guidée  
par ontologie**

Réalisé par :

Lilya Khaleche

Mahamane Massaoudou Ibrahim

Encadré par :

Mme Nora Berrouk

Mme Amel Boustil

Organisme d'accueil :

Centre de recherche sur l'information scientifique et technique (CERIST)

Promotion : 2006/2007

# SOMMAIRE

<b>Introduction Générale</b> .....	1
------------------------------------	---

## Chapitre 1 : «L’Organismes d’accueil »

<b>I. Introduction</b> .....	3
<b>II. Centre de Recherche sur l’Information Scientifique et Technique (CERIST)</b> .....	3
1. Organigramme de CERIST.....	4
2. Service documentation et archives.....	5
2.1 Section ouvrages.....	5
2.2 Section thèses.....	5

## Chapitre 2: « L’Indexation Documentaire»

<b>I. Introduction</b> .....	15
<b>II. Définition</b> .....	15
<b>III. Document</b> .....	15
1. Définition.....	15
2. Documents électroniques.....	15
3. L’acquisition des documents électroniques.....	15
4. Format standard de documents électroniques.....	15
<b>IV. Modes d’indexation</b> .....	16
1. Indexation Manuelle.....	16
1.1 Principe et processus.....	16
1.2 Avantage.....	16
1.3 Inconvénient.....	17
2. Indexation Automatique.....	17
Introduction.....	17
Méthodes d’indexation.....	17
Indexation par mots clés.....	17
Indexation grammaticale.....	17
Indexation morphologico-syntaxico-semantique.....	17

processus d'indexation.....	17
2.3.1 L'acquisition des documents.....	17
2.3.2 L'extraction des mots-clés du document.....	18
2.3.3 La normalisation des mots-clés du document.....	18
2.3.4 L'élimination des mots vides.....	18
2.3.5 La pondération des mots-clés du document.....	18
2.3.5.1 La fréquence d'occurrence.....	19
2.3.5.2 La valeur de discrimination.....	19
2.3.5.3 La $tf*idf$ .....	20
<b>IV. Les fichiers index.....</b>	<b>21</b>
1. Les différents types d'index.....	21
a. Les mots du document.....	21
b. Les Concepts.....	21
c. Les n-grammes.....	22
2. Structures des fichiers index.....	23
2.1. Les fichiers séquentiels (sequential files).....	23
2.2. Les fichiers inversés (Inverted files).....	24
<b>V. Conclusion.....</b>	<b>26</b>

### Chapitre 3 : « les ontologies »

<b>I. Introduction.....</b>	<b>27</b>
<b>II. Définitions.....</b>	<b>27</b>
<b>III. Rôle des ontologies.....</b>	<b>28</b>
<b>IV. Composants des ontologies.....</b>	<b>29</b>
<b>V. Les principaux types d'ontologies.....</b>	<b>29</b>
<b>VI. Langage de spécification d'ontologies.....</b>	<b>30</b>
<b>VII. Construction d'ontologies.....</b>	
1. Les étapes de la construction d'ontologies.....	
2. Outils d'édition d'ontologies.....	
3. Fusion d'ontologies.....	
<b>VIII. Evaluation d'ontologies.....</b>	
<b>IX. Exemples d'ontologies.....</b>	
<b>X. L'Apport des ontologies pour l'indexation documentaire.....</b>	<b>31</b>
<b>XI. Conclusion.....</b>	<b>33</b>

### Chapitre 4 : « Conception d'un système d'indexation de documents PDF utilisant une ontologie OWL »

<b>I. Introduction</b> .....	34
<b>II. Description générale du système</b> .....	34
<b>III. Fonctions du système</b> .....	35
<b>III.1. Extracteur</b> .....	35
<b>III.2. Pondérateur</b> .....	36
<b>III.3. Stockeur</b> .....	36
<b>IV. Conception détaillée du système</b> .....	37
<b>IV. 1. L`extracteur</b> .....	37
<b>a. Extraction des termes</b> .....	38
<b>b. Elimination des mots vides</b> .....	38
<b>c. Lemmatiseur</b> .....	38
<b>IV. 2. Pondération des termes</b> .....	39
<b>1. Calculateur des poids brut des termes</b> .....	39
<b>2. Pondérateur des bloc</b> .....	40
<b>3. Raffineur du calcul du poids des termes</b> .....	41
<b>3.1. L`ontologie</b> .....	41
<b>3.2. L`exploitation de l`ontologie</b> .....	41
<b>IV.3. Fichiers Index</b> .....	42
<b>V. Conclusion</b> .....	44

## **Chapitre 5: «Réalisation et Implémentation du système »**

<b>I. Introduction</b> .....	45
<b>II. Outils utilisés</b> .....	45
<b>II.1. Environnement de développement</b> .....	45
<b>II.2. Outils de manipulation des ontologie</b> .....	45
<b>II.3. Outils d`extraction des termes</b> .....	47
<b>II.4. Outils de lemmatisation des termes</b> .....	47
<b>II.5. Outils de représentation de la base de données</b> .....	47
<b>III. Implémentation des différents modules</b> .....	47
<b>III.1. extracteur</b> .....	47
<b>III.2. pondérateur</b> .....	49
<b>III.3 .construction de l`index</b> .....	50
<b>IV. Architecture du système</b> .....	52
<b>V. Interface du logiciel</b> .....	53
<b>VI. Conclusion</b> .....	55
<b>Conclusion Générale</b> .....	56

**Bibliographie.....58**

**Annexe**

**Annexe A : Source des principales méthodes**

**Annexe B : L'éditeur d'ontologie protégé**

**Annexe C : librairie de JENA**

**Annexe D : Bordereau De Saisie Des Thèses**