

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

Introduction Générale	1
------------------------------------	---

Chapitre 1 : «L’Organismes d’accueil »

I. Introduction	3
II. Centre de Recherche sur l’Information Scientifique et Technique (CERIST)	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»

I. Introduction	15
II. Définition	15
III. Document	15
1. Définition.....	15
2. Documents électroniques.....	15
3. L’acquisition des documents électroniques.....	15
4. Format standard de documents électroniques.....	15
IV. Modes d’indexation	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
IV. Les fichiers index.....	21
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
V. Conclusion.....	26

Chapitre 3 : « les ontologies »

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

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

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

Chapitre 5: «Réalisation et Implémentation du système »

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