

République Algérienne Démocratique et Populaire.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

Université Saad Dahlab, Blida
USDB.

Faculté des sciences.
Département informatique.

**Mémoire pour l'obtention
d'un diplôme d'ingénieur d'état en informatique.**
Option : Système d'information

Sujet :

Indexation des documents structurés en XML

Présenté par : Hammouche Karima
Ouchefoun Houria

Promoteur : Mme Bessai Fatma Zohra
Encadreur : M^r Hadj Yahia Wahid

Organisme d'accueil : CERIST.

devant le jury composé de :

M^{elle} Boustia Narimen

Président

Table des matières

Dédicaces

Remerciement

Sommaire

Table des figures

Introduction générale	2
I.1-Introduction	6
I.2-Définition de W3C	6
I.3-Qu'est-ce que XML?	6
I.3.1-Pourquoi « Extensible » ?	7
I.3.2- Document bien formé et document valide	7
I.3.2.1- Document bien formé	7
I.3.2.2- Document valide	7
I.3.3- Quels sont les Composants XML?	7
I.3.3.1- Elements, Balises [tags]	7
I.3.3.2- Les attributs	8
I.3.3.3- Les commentaires	9
I.3.3.4- L'instruction de traitement	9
I.4- syntaxe du langage XML	10
I.5- structure d'un document XML	12
I.5.1- Un prologue	12
I.5.1.1- Syntaxe de prologue	12
I.5.1.1.1- L'identification XML	12
I.5.1.1.2- L'encodage	13
I.5.1.1.3- La déclaration de (Document Type Definition) DTD	13
I.5.2- Instance de document (l'arbre de l'élément)	14

I.5.2.1- Appels d'entités et sections CDATA	14
I.5.2.1.1- Appels d'entités	14
I.5.2.1.2- La Section CDATA	15
I.6- Définition Type de Document (DTD)	16
I.6.1- La DTD interne	16
I.6.2- La DTD externe	16
I.7- Les espaces de noms	17
I.7.1- Déclaration des espaces de noms	17
I.7.2- Utilisation des espaces de noms	18
I.8- Les feuilles de style	18
I.8.1-Rôle des feuilles de style	18
I.9- Liens XML	19
I.9.1- Le langage Xlink	19
I.9.2- XPointer le langage d'adressage de XML	20
I.10- Analyseur XML (XML parser)	20
I.10.1- DOM (Document Object Model)	21
I.10.2- SAX (Simple API for XML)	22
I.11- Le langage XPATH	24
I.12- XML Schéma	25
I.13- principaux avantages du langage XML	25
I.14- Conclusion	26
ChapitreII : Indexation structurée	
II.1-introduction	28
II.2-La recherche d'information	28
II.2.1- Définition de la Recherche d'Information	28
II.2.2- Les Systèmes de Recherche d'Informations	29
II.3- Documents structurés	33
II.3.1-Définition d'un document structuré	33
II.3.2- A quoi sert un document structuré ?	34
II.4 - La Recherche d'Information Structurée	36

III.2- Schéma XML associé à un livre	53
III.3- Le langage de modélisation unifié (UML)	56
III.4- Démarche de développement du système	58
III.5- Analyse	59
III.5.1- Analyse des besoins	59
III.5.1.1- Les cas d'utilisations	59
III.5.1.2- Diagrammes de cas d'utilisation généraux	61
III.5.2- Diagramme de séquence	70
III.5.2.1- Représentation des interactions	70
III.5.3- Diagramme de collaboration	75
III.5.4- Diagrammes d'activités	77
III.6- Conception	81
III.6.1- Conception générale	81
III.6.1.1- Architecture du prototype Xindexing	81
III.6.1.1.1- Architecture générale du prototype Xindexing	82
III.6.1.1.2- Architecture détaillée du prototype Xindexing	82
III.6.2- conception détaillée	84
III.6.2.1- Le module de lancement d'indexation	85
III.6.2.2- Le module d'indexation du contenu	85
III.6.2.3- Le module d'indexation de la structure	87
III.6.2.4- l'enregistrement des résultats	88
III.7- Conclusion	90
Chapitre 4 : réalisation et mise en œuvre	
VI.1- introduction	92

VI.2- L'environnement de programmation	92
VI.3- prototype Xindexing	92
VI.3.1-Description de l'interface principale du prototype « XIndexing »	93
VI.3.2- Indexation d'un document XML	93
VI.3.2.1-Indexation d'un document en utilisant des paramètres d'importance fixes	94
VI.3.2.1.1- indexation de contenu	97
VI.4.1.1.1- Extraction de contenu de chaque élément	97
VI.4.1.1.2- Elimination des mots vides	98
VI.4.1.1.3- Pondération des termes	100
VI.3.2.1.2-Indexation de structure	105
VI.3.2.1.3-Enregistrement des résultats	108
VI.3.2.2-Indexation d'un document en utilisant des paramètres d'importance variables	108
VI.4-Conclusion	111
Conclusion et perspectives	113
Glossaire	116
Références bibliographiques	132
Annexe A – Les DTD	127
Annexe B – XML Schéma	150
Annexe C – L'API SAX	170