

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediène

Faculté d'Electronique et d'Informatique

Département d'Informatique



Mémoire de fin d'études Présenté
pour l'obtention du diplôme

Ingénieur d'état en informatique

Option : web et technologie XML

Thème :

Génération, par l'exemple,
de modèle de structure
RelaxNG
Pour documents XML.

Proposé et encadré par :

M^r MAREDJ.

Organisme d'accueil : CERIST

Etudié par :

- BOUKHEDOUMA Hocine

- SALAHBRAHIM Abdelhamid

Soutenu le : 05/11/2006

Devant le jury :

M^{me} N.BENSAOU

M^{me} L.MAHDAOUI

M^r L.KADDOURI

Président du jury

membre du jury

membre du jury

Promotion : 2005-2006/n°48

Sommaire

Page

Introduction générale

Chapitre 1 : Technologie XML

Introduction	1
I.Le langage XML	2
I.1 Définition de XML.....	2
I.2 Pourquoi XML ?	2
I.3 Ecriture de documents XML	3
I.3.1 Structure d'un document XML	3
I.3.1.1 Anatomie d'un document XML.....	3
I.3.1.2 Balises, éléments et attributs.....	4
I.3.2 Règles syntaxiques	5
I.3.3 Document bien formé et document valide	5
I.3.4 Arbre XML	6
I.4 Affichage de documents XML.....	7
II.Objectifs officiels de XML	7
III.Quelques domaines d'utilisation de XML	9
IV.Avantages du langage XML	9
V.DTD et W3C XML Schéma	10
V.1 DTD	10
V.2 W3C XML Schéma	10

Chapitre 2 : RelaxNG

Introduction	12
I.Notions de base	12
I.1 XML et la validation	12
I.1.1 Nécessité de la validation.....	12
I.1.2 Les aspects de la validation.....	12
I.2 Schéma RelaxNG	13
I.2.1 Utilisation de RelaxNG pour la validation.....	13
I.2.2 RelaxNG comme un format pivot (format de base).....	13
I.2.3 Dans quels cas utiliser d'autre langages de schéma	14
I.2.4 Vue logique de document.....	14
I.2.5 Les différentes approches des modèles de structure.....	15
I.2.6 Fondement mathématique.....	15
I.2.7 Les motifs	16
I.2.8 Les apports de RelaxNG par rapport au fondement.....	16

II. Les motifs de RelaxNG	17
II.1 Les motifs simples	17
II.1.1 Avec la syntaxe XML	17
II.1.1.1 Le motif <i>text</i>	17
II.1.1.2 Le motif <i>attribute</i>	17
II.1.1.3 Le motif <i>element</i>	17
II.1.1.4 Le motif <i>optional</i>	18
II.1.1.5 Le motif <i>oneOrMore</i>	19
II.1.1.6 Le motif <i>zeroOrMore</i>	19
II.1.2 Avec la syntaxe compacte	20
II.1.2.1 Le motif <i>text</i>	20
II.1.2.2 Le motif <i>attribute</i>	20
II.1.2.3 Le motif <i>element</i>	21
II.1.2.4 Le motif <i>optional</i>	21
II.1.2.5 Le motif <i>oneOrMore</i>	22
II.1.2.6 Le motif <i>zeroOrMore</i>	22
II.1.3 Qu'elle syntaxe doit-on choisir ?	23
II.2 Les motifs nommés	24
II.2.1 Définir un motif nommé.....	24
II.2.2 Référencer un motif nommé.....	24
II.2.3 Les éléments <i>grammar</i> et <i>start</i>	25
II.2.4 Assembler les pièces.....	25
II.2.5 Quelques problèmes évités par RelaxNG.....	28
II.2.6 Modèles Récursifs.....	28
II.3 Les motifs complexes	29
II.3.1 Le motif <i>group</i>	29
II.3.2 Le motif <i>interleave</i>	30
II.3.4 Le motif <i>choice</i>	30
II.3.5 Gestion des espaces blancs.....	31
II.3.6 Contenu mixte.....	32
II.3.7 Combinaison entre les motifs.....	34
III. Imposer des contraintes sur les valeurs des textes	35
III. 1 Le motif <i>value</i>	35
III. 2. L'énumération	36
III. 3 Types de données de RelaxNG (<i>Natives datatypes</i>)	37
III. 3.1 Le datatype <i>token</i>	37
III. 3.2 Le datatype <i>string</i>	38
III. 3.2.1 Utilisation du datatype <i>string</i> dans les valeurs d'attribut.....	38
III. 3.2.2 Doit-on utiliser <i>string</i> ou <i>token</i>	39
III. 4 le motif <i>except</i>	40
III. 5 le motif <i>list</i>	41

IV. Les bibliothèques de datatype	42
IV.1 La bibliothèque Du W3C XML Schéma	43
IV.1.1 Les datatypes	43
IV.1.1.1 Les datatypes string	43
IV.1.1.2 Les datatypes numériques	43
IV.1.1.3 Les noms qualifiés	44
IV.1.1.4 les URIs	44
IV.1.1.5 Les Datatypes <i>binaries</i>	44
IV.1.1.6 Formats de date et d'heure	44
IV.1.2 Les restrictions (facetts)	45
IV.2 La compatibilité des Datatypes de DTD	45
iv.3 Quelle bibliothèque devrait être employée ?	45
IV.3.1 Types natifs contre Datatypes du schéma W3C	45
IV.3.2 DTD contre Datatypes de schéma du W3C.....	46
IV.4 Utilisation de la facette <i>patern</i>	46
IV.4.1 La facette <i>patern</i> comme étant une <i>énumération</i>	46
IV.4.2 Quantification	47
V. Construction de blocs	48
V.1 Utilisation de références externes	48
V.1.1 Avec l'approche 'poupée russe'	48
v. 1.2 Avec la DTD-Like	48
V.2 Fusionner des grammaires	49
V.2.1 Fusionner des grammaires <i>sans redéfinition</i>	49
V.2.2 Fusionner <i>en remplaçant des définitions</i>	49
VI. Les espaces de noms	50
VI.1 L'attribut ns	50
VI.2 Les noms qualifiés	51
VII. Les classes de noms	52
VII.1 La classe <i>anyName</i>	52
VII.2 La classe <i>nsName</i>	52
VII.3 La classe <i>name</i>	53
VIII. Comparaison entre DTD, W3C XML Schéma et Relax NG	54
Conclusion	54
<u>Chapitre 3: Conception</u>	
Introduction	55

I. Problématique et objectifs	55
II. Notre solution	57
II.1 Vue de RelaxNG par rapport au document XML.....	57
II.2 Comment RelaxNG utilise le document XML ?.....	57
III. Description générale du système	58
III.1 Architecture fonctionnelle du système.....	58
III.2 Description de l’algorithme général.....	59
IV. Conception détaillée	59
IV.1 Génération d’un modèle RelaxNG pour le document exemple	59
IV.1.1 Phase1.....	60
IV.1.1.1 Choix du document exemple.....	60
IV.1.1.2 Transformation du document XML en structure type XML.....	60
IV.1.2 phase 2.....	64
IV.1.2.1 Description du processus ‘Norme1’.....	66
A. Traitement des attributs dans Norme1.....	69
B. Traitement des éléments dans Norme1.....	73
B.1 Définition d’une séquence principale.....	73
B.2 Algorithme de traitement des éléments.....	74
B.3 Traitement d’une séquence maximale.....	75
B.4 Génération d’une structure XML Normalisée commune à deux occurrences d’un élément.....	78
IV.1.2.2 Description du processus ‘Norme2’.....	79
IV.1.3 Phase 3.....	81
IV.2 Validation du corpus	83
Conclusion	83
<u>Chapitre 4 : Réalisation</u>	
Introduction	85
I.L’architecture .Net	85
I.1 Les avantages apportés par l’environnement .Net.....	85
I.2 Les couches de l’architecture .Net.....	86
I.3 Le langage C#.....	87
II.Librairies et classes utilisées dans GEMOREX	88
III.Interface du système GEMOREX	89
Conclusion générale	
Annexes A	
Annexes B	

Résumé :

L'utilisation de la technologie XML dans divers domaines d'application (les interfaces avec les bases de données, les systèmes de recherche documentaire, les services web, les parseurs, les générateurs de formulaires XML, etc.) devient de plus en plus effective. La mise en œuvre de ces applications repose essentiellement sur deux préalables : le contrôle rigoureux des données et la connaissance de la structure des documents. La solution réside dans l'utilisation d'un modèle de structure (DTD, Schéma XML ou RelaxNG), qui renseigne, à des degrés divers, sur la nature (type, occurrence, lien hiérarchiques entre constituants, etc.) des informations contenues dans le document XML. De là, il apparaît clairement la nécessité d'accompagner les documents XML par des modèles de structures.

Néanmoins, si nous nous plaçons dans la logique inverse, en d'autres termes, prenons un corpus documentaire XML existant où les documents ne référencent aucun modèle de structure. Là, pour les mêmes raisons citées précédemment, pour pouvoir les exploiter au mieux, il serait intéressant de voir la possibilité de *déduire*, à partir de ces documents, un modèle de structure qui leur assurera la garantie recherchée pour leur traitement.

L'objectif de ce projet est la génération d'un modèle de structure pour documents XML à partir de documents exemples. On s'intéressera en priorité au format RelaxNG qui est un intermédiaire intéressant entre les DTD (simples mais peu expressives) et les Schémas (très expressifs mais très complexes).

Mots clés : XML, DTD, Schéma XML, Schéma RelaxNG, procédé par l'exemple, la validation de documents XML.