



Université Sidi Mohammed Ben Abdellah



Faculté des Sciences et Techniques – Fès

Département de génie électrique
UFR Signaux Systèmes et Télécommunications

THESE DE DOCTORAT

Discipline : Informatique
Spécialité : Informatique
N° d'ordre :

**Indexation et recherche par le contenu des
manuscrits arabes numérisés**

Présentée publiquement par
Noureddine EL MAKHFI

Le 6 avril 2012

Devant le jury composé de :

Président	Pr. Mohcine ZOUAK	PES	Doyen de la Faculté des Sciences et Techniques, Fès
Rapporteurs	Pr. Rachid OULAD HAJ THAMI	PES	ENSIAS, Rabat
	Pr. Hassan QJIDAA	PES	Faculté des Sciences Dhar el Mehraz, Fès
	Pr. Farid ABDI	PES	Faculté des Sciences et Techniques, Fès
Examineur	Pr. Noureddine CHENFOUR	PH	Faculté des Sciences Dhar el Mehraz, Fès
Directeur de Thèse	Pr. Rachid BENSLIMANE	PES	Ecole Supérieure de Technologie, Fès

Table des matières

Remerciements	iii
Résumé.....	iv
Abstract.....	v
Table des matières	vi
Liste des figures.....	ix
Liste des tableaux.....	xii
Liste des abréviations	xiii
Introduction générale.....	1
Partie 1 : Numérisation, exploration et diffusion des manuscrits arabes anciens	
Chapitre I : Structuration et numérisation des manuscrits arabes anciens.....	4
I.1 Patrimoine documentaire au Maroc	4
I.2 Structuration de manuscrits arabes.....	5
I.2.1 Structure physique.....	6
I.2.2 Structure logique.....	8
I.2.3 Structure physico logique.....	10
I.3 Production de documents	10
I.3.1 Production de documents par microfilm.....	10
I.3.2 Production de documents par numérisation.....	11
I.3.2.1 Acquisition des manuscrits.....	12
I.3.2.2 Mise en forme des pages.....	14
I.3.2.3 Prétraitements.....	16
I.3.2.4 Formats et normes de stockage d'images numériques.....	21
I.3.2.4.1 Compression d'images.....	21
I.3.2.4.2 Formats d'images.....	23
I.4 Conclusion.....	25
Bibliographie.....	26
Chapitre II : Métadonnées relatives à la description et à l'indexation des manuscrits numérisés 27	27
II.1 Introduction	27
II.2 Métadonnées.....	27
II.2.1 Métadonnées de Dublin Core :	28
II.2.1.1 Dublin Core simple.....	28
II.2.1.2 Dublin Core qualifié.....	29
II.2.1.3 DCMII (Dublin Core Metadata Initiative).....	30
II.2.2 Métadonnées de l'encodage EAD :	32
II.2.3 Métadonnées de la norme TEI :	32
II.3 Conclusion	37
Bibliographie.....	38
Chapitre III : Encodage et transcription des manuscrits arabes.....	39
III.1 Introduction.....	39
III.2 Transcription des documents numériques.....	40
III.2.1 Formats de documents numériques.....	40
III.2.2 Typologie de formats	40
III.2.3 Passage d'un format d'échange vers un format de présentation	41
III.3 Applications sur les manuscrits arabes	41
III.3.1 Encodage et transcription en HTML	41
III.3.2 Encodage et transcription selon la norme XML TEI.....	43
III.3.2.1 Métadonnées TEI des manuscrits.....	44
III.3.2.2 Structure générale d'un TEI.....	44

III.3.2.3 Expérimentation de la norme XML TEI sur les manuscrits arabes	53
III.4 Conclusion.....	56
Bibliographie.....	57
Chapitre IV : Plateforme d'indexation et de recherche dans les manuscrits arabes.....	58
IV.1 Introduction	58
IV.2 Méthode proposée pour l'accès aux manuscrits.....	60
IV.2.1 Schéma synoptique.....	60
IV.2.2 Schéma architectural de la plateforme proposée	61
IV.2.3 Modélisation de la plateforme proposée	63
IV.2.3.1 Langage et environnement de développement.....	63
IV.2.3.2 Diagramme de classes UML	63
IV.2.3.3 Modélisations XML de la base de données :.....	64
IV.2.3.3.1 Architecture fonctionnelle de la base de données	64
IV.2.3.3.2 Modèle XML simplifié.....	64
IV.2.3.3.3 Modèle XML TEI.....	68
IV.3 Résultats et expérimentations:.....	68
IV.3.1 Présentation de la plateforme proposée	68
IV.3.2 Indexation de nouveaux manuscrits selon des métadonnées.....	69
IV.3.3 Technique de l'annotation des pages de manuscrits	70
IV.3.3.1 Ajout des annotations textuelles de pages	70
IV.3.3.2 Ajout des annotations graphiques de pages.....	73
IV.3.4 Recherche de manuscrits.....	74
IV.3.4.1 Recherche de manuscrits par des métadonnées.....	74
IV.3.4.2 Recherche de manuscrits par des annotations.....	75
IV.3.5 Visualisation	75
IV.3.5.1 Visualisation en mode vignettes	75
IV.3.5.2 Visualisation en mode page par page.....	76
IV.3.5.3 Visualisation en mode tourné les pages.....	77
IV.3.6 Publication de la plateforme	77
IV.3.6.1 Principe.....	77
IV.3.6.2 Déploiement dans un serveur web	78
IV.4 Conclusion.....	79
Bibliographie.....	80
<u>Partie 2: Accès automatique au contenu des manuscrits arabes</u>	
Chapitre V: Segmentation multi-échelle des lignes et des mots des manuscrits arabes.....	82
V.1 Introduction	82
V.2 Sélection de l'espace multi-échelle	84
V.2.1 Sélection de blobs.....	84
V.2.2 Sélection de blobs elliptiques	85
V.3 Méthode proposée	86
V.3.1 Segmentation des lignes dans l'espace multi-échelle	86
V.3.2 Segmentation des mots	95
V.3.3 Résultats de segmentation des lignes et des mots.....	98
V.4 Conclusion.....	100
Bibliographie.....	101
Chapitre VI: Segmentation multi-échelle des caractères imprimés arabes.....	102
VI.1 Introduction.....	102
VI.2 Méthode proposée.....	103
VI.2.1. Binarisation.....	103

VI.2.2 Segmentation des lignes.....	104
VI.2.2.1 Algorithme de la projection.....	105
VI.2.2.2 Histogramme de projection des lignes.....	105
VI.2.2.3 Résultat de segmentation des lignes.....	106
VI.2.3 Segmentation des mots.....	106
VI.2.3.1 Résultats de la segmentation des pseudo-mots.....	106
VI.2.3.2 Résultats de la segmentation des composantes connexes.....	107
VI.2.3.3 Résultats de la segmentation des mots.....	107
VI.2.4 Segmentation des caractères.....	108
VI.4 Conclusion.....	110
Bibliographie.....	111
Chapitre VII: Recherche par le contenu dans les manuscrits arabes.....	112
VII.1 Introduction.....	112
VII.2 Travail relatif.....	113
VII.3 Présentation du système d'accès au contenu des manuscrits.....	114
VII.3.1 Principes de l'approche d'identification de l'écriture manuscrite arabe.....	114
VII.3.1.2 Traitement des manuscrits.....	115
VII.3.1.2.1 Acquisition et prétraitements.....	115
VII.3.1.2.2 Segmentation des objets manuscrits.....	115
VII.3.1.3 Extraction des caractéristiques globales des objets manuscrits.....	116
VII.3.1.3.1 Etat de l'art pour la détection des points d'intérêt.....	116
VII.3.1.3.2 Contribution.....	118
VII.3.1.3.3 Algorithme SIFT «Scale Invariant Feature Transform».....	120
VII.3.1.3.4 Algorithme SURF «Speeded Up Robust Features».....	122
VII.3.1.3.5 Comparaison des points d'intérêt.....	136
VII.4 Expérimentations.....	136
VII.4.1 Détection et comparaison de signatures.....	136
VII.4.1.1 Méthode à base de l'algorithme SIFT.....	136
VII.4.1.1.1 Détection des mots.....	136
VII.4.1.1.2 Détection des zones graphiques.....	136
VII.4.1.2 Méthode à base de l'algorithme SURF.....	138
VII.4.1.2.1 Détection des mots.....	138
VII.4.1.2.2 Détection des zones graphiques.....	139
VII.4.2 Développement des Applications.....	140
VII.4.2.1 Application utilisant la méthode word Spotting.....	140
VII.4.2.2 Application sur la transcription automatique.....	140
VII.4.2.2.1 Ajout des mots inexistantes dans la base de données.....	140
VII.4.2.2.2 Recherche des mots dans la base de données.....	141
VII.4.2.2.3 Algorithme de recherche dans la base de données XML.....	141
VII.4.2.2.4 Conception de la base de données XML.....	142
VII.5 Conclusion.....	145
Bibliographie.....	146
Conclusion générale et perspectives.....	148
Conclusion.....	148
Perspectives.....	149
Bibliographie de l'auteur.....	150
Annexe 1 : Encodage et transcriptions de manuscrits arabes.....	152