

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Normale Supérieure des Sciences Humaines
ALGER

MEMOIRE DE MAGISTER
EN SCIENCES DU LANGAGE ET DE LA COMMUNICATION
LINGUISTIQUE

OPTION : TRAITEMENT AUTOMATIQUE DE LA LANGUE ET IN-
FORMATIQUE LINGUISTIQUE

*PROPOSITION D'UN SUPPORT LOGICIEL POUR
LA GESTION INFORMATISEE DU*
« TRESOR DE LA LANGUE ARABE »

Présenté par :

M. CHERIF-ZAHAR Amine

(Ingénieur d'État en Informatique, option Software - USTHB)

Jury

M. HADJ-SALAH A.,	Président
M. AIT-AOUDIA S.,	Rapporteur de thèse
Mme. DRIAS H.,	Examinateur
Mme. AISSANI A.,	Examinateur

DECEMBRE 2000

Table des matières

INTRODUCTION	11
I. Le Projet du Trfsor de la langue arabe: présentation	11
1. la genèse	11
2- La description:	12
1 - le Trésor: une banque de données textuelles:	12
2 - Le Trésor comme source à la réalisation de divers lexiques	12
3 - le Trésor comme source d'études précieuses	13
2- Fonctionnalités futures du Trfsor:	14
II. Le Trésor: vision informatique	15
III. SGBD, SRI et Systèmes Hypertexts :trois ingrédients pour une bonne gestion de la banque de données du Trfsor	17
III.1. Traitement de l'information dans les « bases de données »	17
111.2. Traitement de l'information dans les systèmes de « Recherche d'Information »	17
111.3. Traitement de l'information dans les systèmes de « navigation hypertextuelle »	18
III.4. Comparaison des trois procédés de traitement de l'information	18
III.5. Impact des trois procédés sur les travaux de réalisation et de gestion du Trésor	19
IV. Le système proposé:	19
LE SYSTÈME DE GESTION DE BASE DE DONNÉES DU TRÉSOR	23
ASPECT MODÈLES DE DONNÉES ET MODÈLES DE REPRÉSENTATION	26
III.1. Introduction	26
III.2. Le modèle de représentation des données	28
III.2.1. Bref état de l'art des « modèles de données » utilisés en « base de données »	29
II.2.1.1 modèles généraux [Del, 90]	29
11.2.1.1.1. Les modèles sémantiques	29
III.2.1.1.2. Les modèles de représentation d'objets structurés	31
111.2.1.13. Les bases de données « orienté-objet » [Ami, 92][Del, 90]	32
III.2.2. Interférences entre les différentes classes de modèles généraux	33

III.2.3. modèles de représentation de données textuelles	33
III.3 Le « monde réel » à modéliser	35
111.4. Choix du modèle	36
111.5. Description du modèle choisi : « Le modèle d'objets complexes »	37
III.5.1. Les objets complexes [Abi, 87], [Val, 87], [Del, 90]	38
III.5.2. Description du modèle [Abi, 87]	38
III.5.3. Identité d'objet [Val, 87], [Del, 90]	39
III.5.4. Techniques de représentation des identifiants d'objet [Del, 90]	40
III.5.4.1. Adresse physique	40
III.5.4.2. Indirection	40
III.5.4.3. Identificateur interne	40
III.5.4.4. Clé	41
III.5.4.4. Identificateur structuré	41
III.5.5. Représentation des données du système par le modèle retenu	41
111.5.1.1. Modèle direct	41
111.5.1.ZModèle normalisé	42
III.5.1.2.1. Variantes du modèle normalisé	43
IV. Niveau conceptuel : Modélisation de la réalité textuelle par le modèle d'objets complexes	43
LES ASPECTS: SYSTÈME DE STOCKAGE PHYSIQUE DES DONNÉES ET PRIMITIVES DE BAS NIVEAU	47
II.1. Introduction	47
II.2. Exposé des classes de systèmes de stockage	47
II.2.1. Systèmes de stockage et méthodes d'accès des systèmes relationnels [Del, 82]	47
a) L'organisation en tas	47
b) Le hachage ou adressage associatif	47
c) Les fichiers indexés	48
c.1) Les indexes creux	48
c.2) Les indexes denses	49
II.2.2. Représentation des données en MC et MS par les gestionnaires d'objets	50
II.2.2.1. Avantages et inconvénients des deux méthodes	51
II.2.2.2. Systèmes à un niveau d'adressage : l'exemple de Socrate	51
II.2.2.3. Système à deux niveaux d'adresses: l'exemple de PS-Algol	52

11.3 Techniques de gestion des données dans les SGBD	52
II.3.1. Représentation normalisée d'une structure	53
II.3.1.1. Représentation des ensembles	54
II.3.2. Représentation linéaire	55
II.3.2.1. représentation des champs courts et longs	55
a) représentation des objets courts:	55
b) représentation des objets longs	56
c) représentation des grands ensembles et des chaînes longues	57
II. 4. Description du système de stockage adopté	59
II.4. 1. Discussions de techniques d'organisation	59
II.4.1.1. Discussion de cette dernière stratégie:	61
II.4.1.2. Amélioration de la précédente technique:	62
II.4.1.3. Autre solution:	63
II.4.2. Routines de gestion des deux espaces:	65
II.4.2.1. gestion du premier niveau de représentation	65
II.4.2.2. gestion du second niveau de représentation	65
II.4.3. Mise en œuvre d'un support d'identité d'objet dans notre système	65
II.4.4. Choix final de l'organisation	66
II.4.5. Description formelle: structures de données et routines de manipulation:	66
II.4.5.1 - représentation de la structure d'une base de données:	66
II.4.5.2 - Les index (relationnels) pour les besoins de la recherche d'attributs structurels	69
II.4.5.3. L'espace des champs courts:	70
II.4.5.4. L'espace des champs longs:	70
II.4.5.5. Les migration d'objets entre la mémoire centrale et la mémoire secondaire.	70
II.5. Implémentation effective	71
II.6. Conclusion	71

LES ASPECTS LANGAGE DE DESCRIPTION DES DONNÉES ET LANGAGE DE REQUÊTES

73

111.1. Le langage de définition des données	73
III.1.1. Syntaxe du LDD:	73
III.1.2. Plan de traduction d'une définition	77
II.1.3. Exécution de la phase de traduction:	78
II.1.4. Opération d'assimilation et d'extraction de d'instances:	79
111.2. Le langage de requêtes	81

III.2.1. Le langage: Aspect traitements et apports de l'algèbre d'objets complexes	82
III.2.1.1. Nature des requêtes:	82
III.2.2. Exposé de quelques systèmes et de leurs langages de requêtes	83
111.2.2.1. Algèbres et modèles de données	83
1 - l'algèbre relationnelle	84
2- l'algèbre des modèles non en première forme normale:	84
3- l'algèbre de Gütting (pour documents structurés)	90
5- l'Algèbre d'Objets complexes:	93
III.2.3. description informelle du langage adopté	94
III.2.4. Génération de plans d'exécution	101
111.2.4.1 - Représentation des composantes intentionnelle et extensionnelle	101
III.3. Conclusion	107

LE SYSTEME DE RECHERCHE D'INFORMATIONS ET LA NAVIGATION DANS LE TRÉSOR **108**

LE MODULE SYSTEME DE RECHERCHE D'INFORMATIONS **109**

IV.1. Techniques de recherche d'Information	109
IV.2. Proposition d'une amélioration de la technique des fichiers inverses:	111
IV.2.1. Discussion de la méthode:	112
IV.2.2. Améliorations de la précédente technique	112
IV.3. Proposition d'un système d'indexation basé sur une décomposition morphologique des mots	113
IV.3.1. Caractéristiques morphologiques de l'arabe [Coe, 93]	113
IV.3.1.1. Structure du lexème	113
a) La racine	113
b) Le schème	114
IV.3.2. Algorithme de décomposition morpho-lexicale [PFE, 95]:	116
IV.3.3. Un autre algorithme: [Tai, 97]	118
IV.4. Éléments nécessaires à une bonne décomposition: les schèmes des lexies nominale et verbales	118
IV.4.1. La lexie nominale:	119
IV.4.2. la lexie verbale:	119
a) le schème du verbe à l'accompli:	120
b- schème du verbe à l'inaccompli	121
c - lexie du verbe à l'impératif	122

IV.5. Proposition d'un algorithme de décomposition morphologique basé sur la notion de lexie 123

IV.6. Évaluation qualitative et quantitative du système 124

LES MODULE DE NAVIGATION « HYPERTEXTE » 125

V.1. Les systèmes « Hypertexte »: notions générales 125

V.1.1. Définitions: [Gin, 93] 126

V.1.1.1. Hypertexte. 126

V.1.1.2. Nœuds et liens 126

V.1.1.3. Les modes de lecture 127

V.1.1.4. Problèmes liés à la navigation 127

V.1.1.5. Quelques applications et quelques systèmes 128

V.1.1.6. Schéma descriptif d'une portion d'un hyperdocument 128

V.1.2. Les composantes d'un système Hypertexte 128

V.2. Passage d'un document linéaire à un hyperdocument et inversement: 129

V.2.1. Hyperdocument vers document linéaire 130

V.2.1.1. la méthode manuelle: 130

V.2.1.2. la méthode semi-automatique: 130

V.2.1.2.1. Règles à observer pour faire migrer un document depuis sa forme linéaire à une forme Hypertexte: 130

V.2.1.3. La méthode entièrement automatique 131

V.3. Proposition d'un module de gestion de liens « hypertextuels » entre données 131

V.3.1. les idées 131

V.3.2. la conception 132

V.3.2.1. le gestionnaire de nœuds et de liens: 132

V.3.2.1.1. Le réseau de premier niveau: l'indexe général du Trésor: 132

V.3.2.1.2. génération du réseau de liens: 133

V.3.2.2. le créateur de liens: 133

V.3.2.3. l'interface graphique 134

V.4. Discussion des avantages et des inconvénients du module en question et les améliorations possibles du système 134

LE TRÉSOR: UNE ARCHITECTURE DISTRIBUÉE SUR LE RÉSEAU INTERNET 135

VI.1. Quel intérêt offre Internet dans le cadre d'un traitement réparti des données ? 135

VI.2. Quel type de réseau adopter?	136
Solutions aux problèmes posés	139
VI.3. Architecture du système « distribué »	139
VI.4. Conclusion sur l'architecture distribuée:	142
CONCLUSION GENERALE	144
BIBLIOGRAPHIE	146