

MINISTERE DE L' EDUCATION NATIONALE

SECRETARIAT D' ETAT A LA RECHERCHE

**THESE**

Présentée au

CENTRE DE DEVELOPPEMENT DES TECHNOLOGIES AVANCEES

Pour l'obtention du grade de

**MAGISTER EN CYBERNETIQUE**

( Option : Architecture des systèmes )

**OPTIMISATION DU STOCKAGE ET DE L' ACCES  
DANS LES BASES DE DONNEES  
TRES VOLUMINEUSES**

Par

**Mme Fatma-Zohra BESSAI  
née MECHMACHE**

Soutenue le 45 Novanbre 1992 Devant le jury composé de

Mr H. Bessalah	.....	Président
Mr A. Abdellaoui	.....	Examineur
Mme Z. Ali-Mazighi	.....	Examineur
Mr M. Benhamadi	.....	Examineur
Mme F. Cherief	.....	Examineur
Melle F. Azrou	.....	Rapporteur

## Remerciements

Je tiens a remercier,

Monsieur H. BESSALAH, Directeur du centre de developpement des technologies avancees (CDTA), pour avoir bien voulu me faire l'honneur de presider le jury.

Monsieur M. BENHAMADI, Directeur du CERIST, pour avoir mis a ma disposition les moyens necessaires a la bonne marche de mon travail, pour la confiance qu'il m'a temoignee, et pour sa participation au jury.

Je suis tres honoree de la participation a ce jury de Madame Z. ALI-MAZIGHI, chargee de cours a l'USTHB, Madame F. CHERIEF, maitre de conference (USTHB) et de Monsieur A. ABDELLAOUI, maitre de conference (USTHB).

Je tiens egalement a remercier Mademoiselle F. AZROU, chargee de recherche au CERIST, pour avoir propose le sujet de recherche et dirige mon travail.

Je ne peux pas oublier ici H. LABIOD, Ingenieur au CERIST, pour son aide et sa collaboration pendant l'integration de mon travail au systeme SIBADOC.

Je ne saurais clore ces pages sans adresser mon expression de sympathie la plus sincere a tous mes collegues et amis du Laboratoire de Recherche et Developpement en Informatique (LRDI), ainsi qu'a l'ensemble du personnel du CERIST. Que toutes celles et ceux qui m'ont temoigne leur disponibilite et apporte leur aide trouvent ici une expression de reconnaissance.

Enfin je tiens a remercier tous mes proches pour leur soutien et patience tout au long de l'elaboration de ce travail. Qu'ils trouvent ici une recompense a leurs efforts.

## RESUME

La masse d'information mise a la disposition de la communauté scientifique s'accroît de façon continue, étant donnée l'évolution et la diversification des disciplines scientifiques et techniques. Cette grande masse d'information nécessite l'optimisation des moyens de stockage et le développement de nouvelles méthodes d'accès car les méthodes classiques se révèlent insuffisantes.

L'objectif de notre travail est de contribuer à l'amélioration des techniques d'accès aux bases de données volumineuses. Or une méthode d'accès est étroitement liée à la méthode de stockage utilisée. Aussi, une partie de notre travail a consisté à étudier le compactage de données, utilisé au CERIST pour résoudre le problème de stockage de grands volumes de données.

Après l'étude des diverses techniques de compression, notre travail a essentiellement consisté à concevoir et à réaliser un système d'accès aux données volumineuses. Ce système utilise une technique d'indexation basée sur l'inversion totale, ce qui permet aux utilisateurs d'accéder à la base de données par n'importe quel mot contenu dans cette dernière et offre un accès rapide à l'information grâce à l'utilisation du Q-arbre qui est un arbre équilibré de profondeur 1. En plus de ces caractéristiques, le système a une structure modulaire lui permettant de s'adapter à tous les systèmes gérant de grands volumes de données textuelles.

## MOTS CLES

Base de données documentaire, système de recherche documentaire, système de gestion d'objets complexes, stockage des données, accès aux bases de données, compression de données, indexation, Q-arbre.

## **ABSTRACT**

The amount of scientific information is growing steadily. This huge amount of information requires efficient storage management and new access methods.

Our goal is to improve the current access techniques to very large data bases. Since an access technique is closely related to the storage method used, in the first part of our work we study the data compression techniques already used by the CERIST in order to store a large amount of information.

After an analysis of the current techniques of data compression, we concentrate on the design and implementation of an access system to very large data bases. Our system uses an indexing scheme based on full inversion which allows users to query the data base using any word contained in the data base. This system is based on the Q-tree (B-tree of depth one) that indeed allow fast access to the sought information. In addition, the system modular structure makes it adaptable to all system managing very large textual data.

## **KEY WORDS**

Bibliographic data base, document-retrieval system, complex object management system, data storage, data base access, data compression, indexing, Q-tree.

## S O M M A I R E

Introduction

I	Evolution des systemes de gestion de donnees	
1.1	Les systemes de recherche documentaire .....	5
1.2	Les systemes de gestion de bases de donnees .....	7
II	Compression de donnees	
11.1-	Definition .....	13
11.2-	Concepts de base de la compression de donnees (CD) .	13
11.2.1-	Le code et ses proprietes .....	14
11.2.2-	Quantité d'information et optimalité .....	16
	d'un code	
11.2.3-	Schema de compression de donnees .....	17
11.2.4-	Parametre d'évaluation des techniques .....	19
	de compression de donnees (TCD)	
11.3-	Les redondances dans la représentation des donnees .	19
11.4-	Classification des techniques de compression de ....	20
	donnees	
11.5-	Choix d'une TCD pour la compression des donnees ....	25
	bibliographiques	
11.6-	Integration de la technique de compression de .....	29
	donnees au systeme SIBADOC	
III	Techniques d'accès	
111.1	Description des differentes techniques d'accès ...	37
III.1.1	Recherche sequentielle .....	37
	(Full Text Scanning)	
III.1.2	Inversion .....	37
III.1.3	Regroupement (Clustering) .....	38
III.1.4	Fichier signature (Signature file) .....	39
III.1.5	Methode d'accès utilisant un fichier ....	40
	signature	
III.1.6	Methode d'accès utilisant un Q-arbre ....	43
III.1.6.1	Definition d'un Q-arbre .....	43



v.2.2 Module d'accès aux donnees ..... 71  
v.2.3 Module de mise a jour ..... 73  
v.3 Evaluation du systeme d'accès aux donnees ..... 73  
volumineuses  
v.3.1 Evaluation de l'indexation ..... 73  
v.3.2 Evaluation de la methode d'accès ..... 77

Conclusion et perspectives ..... 80

Bibliographie

Annexe A: SIBADOC : Un systeme d'interrogation des bases de  
donnees documentaires

Annexe B: Liste des mots vides utilisés par le module  
d'indexation du systeme d'accès aux donnees  
volumineuses.