Système d'Aide à la Conception de Bases de Données en Langage Naturel(*)

O. Nouali

Laboratoire de Recherche et Développement en Informatique,
CE.R.I.S.T.
E-mail : Nouali@ist.cerist.dz

Abstract

This paper présent a database design System in natural language.The main features of the System is : a natural language interface and expert System. With interface, a user may specify databases in natural language.The expert System converse with the user,detecte incoherence,make a complète database conceptual scheme and finally provide a relationnel scheme.

Introduction

La complexité et la taille sans cesse croissantes des applications bases de données(B.D) nécessitent le développement de méthodes systématiques facilitant la conception de ces bases en assistant leurs concepteurs.

Certaines parties du processus de conception ont pu être formalisées à l'aide d'algorithmes: dans ce cas si l'algorithme est complexe, des programmes qui l'automatisent sont développés pour alléger le travail des concepteurs. D'autres parties du processus sont de nature purement heuristique. On ne peut alors raisonnablement qu'assister le concepteur dans sa tâche, en l'aidant à définir ses objectifs (en particulier, les étapes ou modèles définis par la méthode) en lui fournissant des formalismes adaptés à son problème, en lui suggérant des règles ou des indications, et enfin, en lui offrant **des** programmes de conception **assistée par** ordinateur (CAO)[BOU 85].

Pour ne pas astreindre l'utilisateur à l'apprentissage d'un langage technique, le processus de conception serait grandement simplifié par une **spécification en L.N.** des **besoins en données d'une** application. **Cependant,** une **difficulté** concerne la grande variété de **styles** de phrases que **supporte un L.N.** Il **apparaît** que l'on doit autoriser seulement les styles qui conduisent à une modélisation directe (et, de préférence, unique) des besoins d'une application[BAT 92].

La nature de la tâche de conception exige la mise en œuvre de deux types de connaissances (connaissances sur le(s) modèle(s) conceptuels et connaissances expérimentales). La construction d'un outil qui reproduise l'attitude de l'expert et qui exploite une connaissance expérimentale, indispensable à la maîtrise de la conception en la combinant à une connaissance plus formelle, peut apporter une aide "réelle" dans le processus de conception.

(*) Communication faite lors du Deuxième Symposium Internationsl sur la Programmation et les Systèmes/ ENSAG, Alger, 10-12 avril 1995.

Le système a été construit dans le but de répondre aux objectifs suivants :

(1) Faciliter l'interaction avec l'utilisateur en lui offrant une interface conviviale aussi riche que facile à utiliser Cette interface doit notamment lui permettre de capturer et d'exprimer toute la sémantique de son application.

(2) Construire des bases de connaissances regroupant des acquis théoriques sur les modèles tels que le modèle Entité/Association(E/A) et le modèle relationnel.

(3) Identifier pour chaque phase de conception les mécanismes de raisonnement généraux ou spécifiques, empiriques ou algorithmiques.

(4) Construire un système d'outils ouvert capable d'acquérir de nouveaux concepts théoriques et de nouvelles règles de conception.

Dans cet article, nous nous proposons tout d'abord de donner le fonctionnement général et l'architecture du système. Nous décrivons ensuite l'interface en L.N. et la méthodologie de conception de bases de données. Nous terminerons par un exemple de session.

1. Fonctionnement général du système

Le système d'aide à la conception de B.D. en L.N. est composé d'une interface en L.N., d'un traducteur et d'un S.E.
L'interface analyse la spécification énoncée par l'utilisateur; puis produit une représentation interne du sens de cette dernière.
Le traducteur transforme la représentation interne produite en une base de faits représentant le schéma conceptuel modélisant les données.

Cette base de faits est ensuite interprétée par un système expert que nous avons spécialisé dans la conception de B.D.

Premièrement, le système expert vérifie la cohérence (absence de contradictions et de redondances) et la complétude (présence de toutes les informations nécessaires) de la base de faits considérée.
En cas d'échec (présence de contradictions ou absence d'informations nécessaires à la poursuite du processus . de conception), le concepteur est invité à corriger sa spécification en fonction des diagnostics qui lui sont fournis.
Le processus de conception est en effet un cycle d'acquisition/validation [MET 93] qui ne s'arrête que lorsque la spécification est jugée correcte par le système expert. •

Deuxièment, le schéma conceptuel, validé et complété est traduit en un schéma logique relationnel. Ce schéma est produit indépendamment du S.G.B.D. qui sera effectivement utilisé par la suite pour créer la base de données.
La nouvelle base de faits produite (contenant le schéma relationnel) sera traduite en une spécification textuelle codée dans un langage de description de données (L.D.D. e.g. SQL) correspondant au SGBD utilisé pour créer la B.D. conçue.

2. Architecture du système

(voir schéma page suivante)

3. l'interface en langage naturel

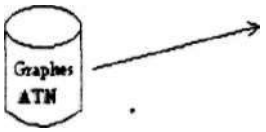
L'interface en L.N. représente le niveau linguistique du système : Elle est destinée à faciliter la tâche du niveau conceptuel pour modéliser une B.D. Pour des raisons de modularité et d'efficacité, elle est divisée en deux modules :
- un analyseur "Morpho-lexical"
- un analyseur "Syntaxico-sémantique".

Saisie
langage I naturel

(spécification
texte externe)

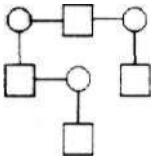
dut
(mots*
concepts

(Lexique îenuntujue (fan domaine
dappbcat »n choisi)



ANALYSEUR
DU LANGAGE NATUREL

erreurs diagnostic:



(représentation interne
en graphes conceptuels
de Sowa)

traducteur

(Passage Graphes concept™ '•
de SOWA —> schéma conceptuel E/A)

(schéma conceptuel
BFi version non validée)

cohérence |
+
complétude j

BRI
vahdatioi

erreurs diagnostic!

GENERATEUR
DE
SYSTEMES EXPERTS

schéma
conceptuel

BR2

(passage schéma conceptuel
» schéma relationnel)

(CERESY)

schéma
relationnel

BR3
LDD "X

(règks pour LDD du SGBD "X")

schéma
physique

(Spécification du schéma physique
relationnel dans le LDD du SGBD "X")

u

(schéma conceptuel
version validée) BF_2 BF_3 **(schéma relationnel**
indépendant de tout SGBD)

3.1. L'analyseur morpho-lexical [NOU 91], [AZO 94]

L'analyseur "morpho-lexical" reçoit en entrée une phrase écrite en langage naturel, la découpe en entités lexicales et restitue en sortie la ou les définitions de chacune d'elles. Un lexique est constitué d'un noyau de base qui est augmenté, pour chaque application, des mots spécifiques à cette dernière. Ce noyau regroupe les mots-outils(articles, prépositions, pronoms,...) et les mots propres à la conception de bases de données(e.g. verbes exprimant des dépendances fonctionnelles, des généralisations, ...).
La méthode consiste à ne stocker dans le dictionnaire que les formes canoniques (singuliers lorsqu'ils existent, bases des verbes, ...), ce qui réduit considérablement le volume du dictionnaire. En contre-partie, ' elle nécessite des mécanismes d'analyse permettant de passer d'une forme fléchie à une forme canonique et des mécanismes de recherche pour accélérer les accès.

3.2. L'analyseur **syntaxico-sémantique[NOU 91], [AZO 94]**

L'analyseur "syntaxico-sémantique" reçoit en entrée une liste de mots dotés de leurs définitions et restitue en sortie le sens de la phrase.
L'étude des différents outils existants nous a menés à opter pour l'utilisation des grammaires ATN(Augmented Transaction Network) sémantiques pour la réalisation de l'analyseur syntaxico-sémantique. Elles permettent de mettre en relief l'interdépendance de la syntaxe et de la sémantique en permettant une analyse dirigée par la syntaxe tout en construisant, au fur et à mesure de l'analyse, des parties de représentation interne du sens grâce aux actions attachées aux arcs de la grammaire. En faisant le bon choix des structures syntaxiques à représenter par les graphes de la grammaire et le jeu des conditions sélectionnant les chemins à emprunter, l'analyse peut être très efficace étant donné que le nombre de concepts est relativement réduit.
La grammaire ATN se compose d'un ensemble de graphes implémentés sur une structure de données dynamique qui permet facilement de l'enrichir. Le parcours de ces graphes, pour extraire le sens de la phrase, est réalisé par un interpréteur ATN. Pour des raisons de flexibilité et de portabilité, il a été conçu de manière à ce qu'il soit général et indépendant de tout domaine. Pour considérer un domaine particulier, il suffira de disposer du dictionnaire et de la grammaire ATN correspondants. La seule contrainte à respecter étant le codage de la grammaire.

3.3. Représentation de sens

En informatique, comprendre une phrase écrite en langage naturel consiste à transformer celle-ci en une représentation interne manipulable par programme. Il s'agit donc ici de trouver une structure de données permettant de représenter le sens d'une phrase.
Les graphes conceptuels sont un modèle adéquat pour la représentation interne du sens de faits exprimés en langage naturel [MET 93], [MOU 92].
Chaque élément de connaissance est représenté par un graphe fini, biparti, orienté comportant deux types de noeuds : des *concepts* et des *relations* liant des concepts. Dans un graphe, les noeuds "concept" représentant des entités, des attributs, des états, et des événements; et les noeuds "relation" montrent comment les concepts sont interconnectés [FAR **89].**

Les graphes conceptuels sont aussi un outil d'analyse sémantique grâce à des primitives de manipulation de graphes, principalement l'appariement et la jointure des graphes conceptuels [CHE 92].
L'analyseur détermine le sens de la phrase en réunissant les graphes conceptuels de tous les éléments de la phrase, en s'aidant de la structure syntaxique de la phrase. Il utilise comme support de base un lexique sémantique contenant les graphes canoniques nécessaires pour coder le sens des mots. Pour chaque mot de la phrase, le programme choisit un graphe conceptuel parmi les différentes alternatives possibles : il examine un sommet de l'arbre syntaxique (pouvant être implicite) couvrant plusieurs mots et effectue l'opération d'appariement des graphes. Cette opération échoue quand il existe une incompatibilité entre certains concepts intervenant dans les graphes des deux mots que l'on tente d'associer; dans un tel cas, le programme cherche d'autres graphes conceptuels correspondant à d'autres significations possibles des mots incompatibles, et il tente à nouveau l'appariement. Finalement, l'analyse produit un graphe conceptuel représentant la signification du groupe de mots situé sous le sommet considéré de l'arbre syntaxique. De proche en proche, et en remontant vers le sommet de l'arbre syntaxique, le programme construit un graphe conceptuel représentant la signification de la phrase.

L'analyse sémantique utilisée par notre système est dirigée par la syntaxe grâce à une grammaire ATN [AZO 94]. Le codage du lexique sémantique est une tâche difficile, et la qualité de ce codage conditionne le résultat global du projet d'analyse sémantique.

4. Méthodologie de conception de bases de données [AZO 94]

Le processus de conception est long et complexe. Il débute par l'analyse des besoins d'un utilisateur en données et aboutit à un schéma physique relationnel. Il est alors nécessaire de suivre une méthodologie rigoureuse permettant d'appréhender les différents aspects de ce processus de conception.
La méthodologie de conception de bases de données adoptée est basée, au niveau conceptuel, sur un modèle entité-association étendu. En effet, il est souvent utilisé comme outil de communication entre un concepteur et un utilisateur de la base de données, grâce à sa facilité d'utilisation et de représentation des concepts [STO 91].

4.1. Le modèle Entité-Association étendu

Les principaux concepts de ce modèle sont les concepts d'entité et d'association entre les entités (chacune jouant un rôle particulier). Les entités et les associations peuvent être caractérisées par des attributs représentant des propriétés significatives. Les liens entre entité et association sont caractérisés par un couple : cardinalite minimale et cardinalite maximale. Toute entité doit être identifiée de façon unique par un ou plusieurs de ses attributs et/ou d'autres entités. Des dépendances fonctionnelles existent entre des attributs. Les hiérarchies de généralisation permettent d'avoir des schémas conceptuels plus concis et une sémantique plus précise grâce au mécanisme d'héritage.

4.2. Description de la méthodologie de conception

La méthodologie de conception consiste

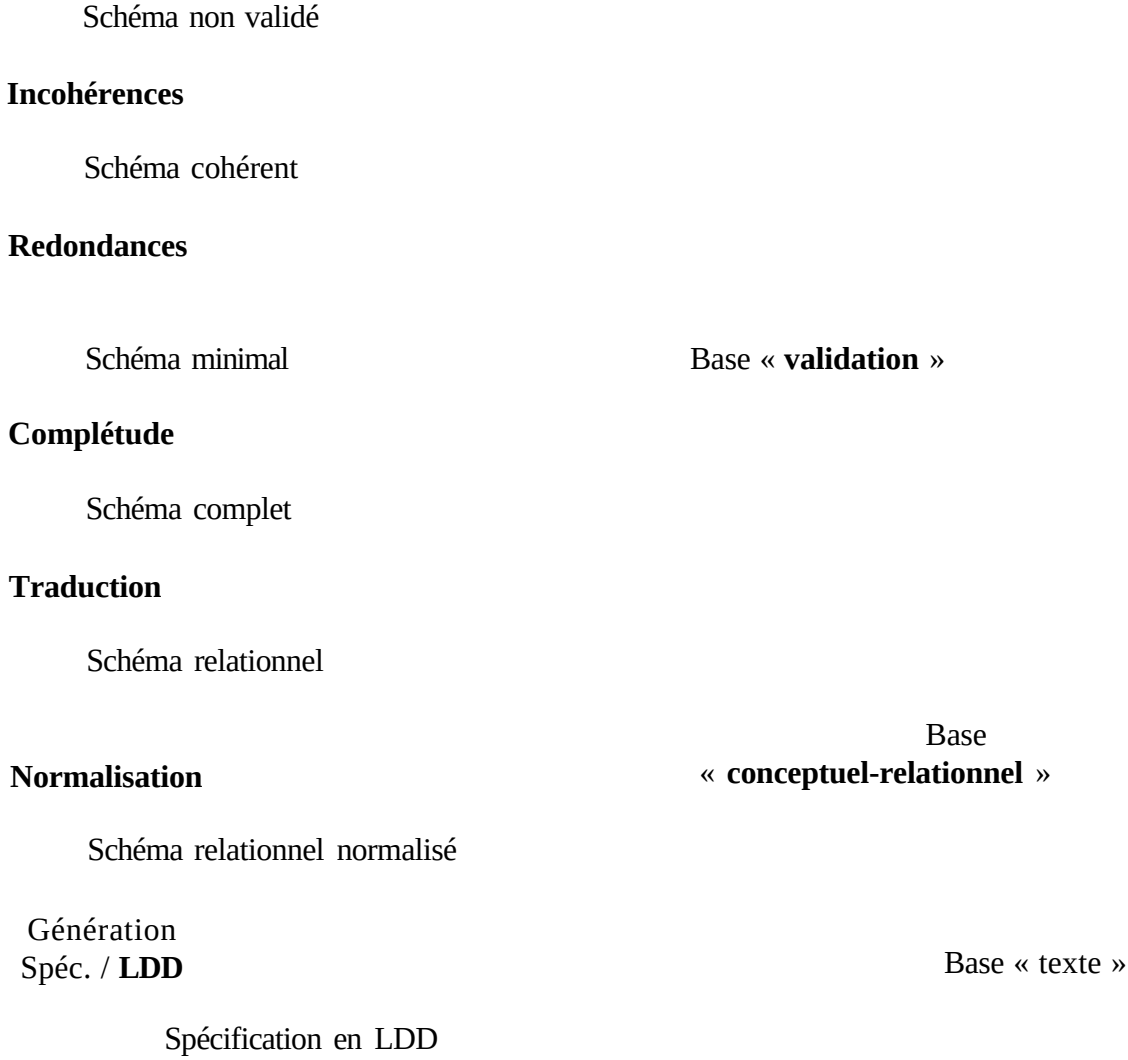


Figure 1 : Méthodologie de conception adoptée

en des règles d'inference (sous forme de règles de production) destinées à être interprétées par un générateur de systèmes experts. Ces règles sont réparties sur plusieurs bases de règles: « validation », « conceptuel-relationnel » et « texte » (figure 1).

Deux types de règles sont distingués :

des méta-règles (règles fixant l'application éventuelle et l'ordre d'application d'autres règles) qui pilotent, de façon globale, le processus de conception par enchaînement des différentes étapes,

- des règles qui concourent à la réalisation de tâches particulières.

L'utilisation de méta-règles offre l'avantage d'ajouter facilement de nouvelles étapes au processus de conception au fur et à mesure que la stratégie est raffinée (par l'ajout de règles suggérées par des experts).

La base "validation" consiste à vérifier :

- La cohérence :

*détecter et signaler les incohérences du schéma conceptuel initial correspondant

à la spécification fournie par l'utilisateur.
* détecter les redondances et. le cas échéant, rendre le schéma conceptuel minimal,
- La complétude détecter les informations manquantes et rendre le schéma conceptuel complet conformément au modèle conceptuel choisi,
- Générer et afficher le schéma conceptuel
obtenu pour être validé par le concepteur qui vérifie la
conformité du schéma avec sa spécification.

La base "conceptuel-relationnel" consiste à :
- choisir un identifiant pour chaque entité parmi ses candidats possibles suivant certains critères,
- traduire le schéma conceptuel en un schéma relationnel avec les contraintes d'intéqrité
nécessaires,
- normaliser les relations obtenues à létape précédente pour obtenir un schéma relationnel normalisé (dans notre cas, nous nous limitons à
la troisième forme normale),
- afficher le schéma relationnel obtenu.
sb sm îneq -
La base "texte" consiste à générer la spécification du schéma relationnel obtenu dans un langage de description de données correspondant à un SGBD particulier.

5. Exemple d'une session de travail

L'exemple décrit ci-dessous concerne la gestion de bibliothèques.

Spécification en L.N.

Un livre et un article sont une sorte d'ouvrages.

Un lecteur peut emprunter un ou plusieurs ouvrages pour une certaine période.
Les ouvrages sont décrits par un résumé et des mots-clés.
Un livre est identifié par un ISBN et un numéro d'édition

Représentation en graphes conceptuels

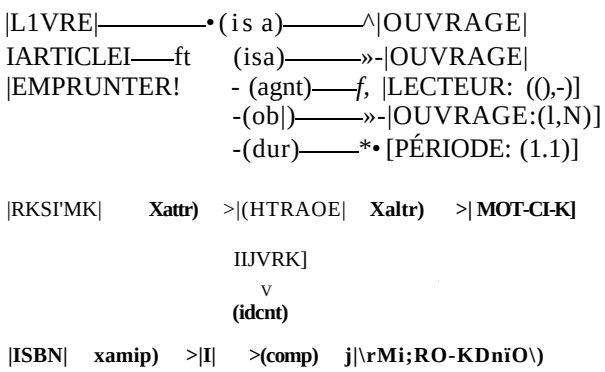


Schéma conceptuel E/A (base **de faits)**

(GEN LIVRE OUVRAGE)
(ENTITE OUVRAGE)
(ENTITE LIVRE)
(GEN ARTICLE OUVRAGE)
(ENTITE ARTICLE)
(AS SOC 1 EMPRUNTER)
(E-R 1 LECTEUR AGNT 0 N)
(E-R 1 OUVRAGE OBJ 0 1)
(ATT-ASS 1 PERIODE I 1)
(ENTITE LECTEUR)
(ATT-ENT RESUME OUVRAGE 1 1)
(ATT-ENT MOT-CLE OUVRAGE 1 N)
(ID-ATT 1 LIVRE ISBN)
(ID-ATT 1 LIVRE NUMERO-EDITION)
(ID-ATT 2 OUVRAGE COTE)
9QSÇn6l

Schéma relationnel (base de faits)

...
(REL LECTEUR)
(KEY LECTEUR NUMERO-LECTEUR)
(REL OUVRAGE)
(KEY OUVRAGE COTE)
(REL EMPRUNTER)
(KEY EMPRUNTER COTE)
(FORE.IGN EMPRUNTER LECTEUR)
(ATT EMPRUNTER PERIODE 1 1)

Spécification en SQL

```
CREATE TABLE EMPRUNTER!
        COTENUMBER(5).
        NUMERO-LECTEUR NUMBER (4).
        PERIODE DATE NOT NUEE UNIQUE.
        PRIMARY KEY (COTE) CONSTRAINT
        INCLOUVRAGE.
        FORE1GN KEY (NUMERO-LECTEUR)
                REFERENCES LECTEUR (NUMERO-
                LECTEUR)
```

Conclusion

Le Système d'aide à la conception de bases de données en langage naturel réalisé nous a permis de montrer l'intérêt et la faisabilité de l'approche système expert. Il nous a également montré quelques problèmes importants dans la construction d'une interface en langage naturel (l'énumération des connaissances nécessaires à la compréhension, les théories formelles qui permettent de les représenter et les mécanismes informatiques pour les utiliser) et dans la formalisation de la connaissance experte et sa traduction en prédicats et en règles de production.

Le système se compose de deux parties principales :

une interface analysant des spécifications en langage naturel énoncées par l'utilisateur modélisant une base de données,

- un système expert pilotant le processus de conception.

La liaison entre ces deux parties est assurée par un traducteur qui s'occupe de traduire une représentation interne du sens des phrases fournies en prédicats interprétables par le système expert.

Le formalisme de représentation **interne** du sens est le modèle des **graphes** conceptuels de Sowa [SOW 86]. La puissance d'expression sémantique de **ce** modèle permet d'appréhender tous **les** aspects utiles à la modélisation conceptuelle d'une base de données.

La Méthodologie de conception est basée sur un modèle E/A étendu. Elle est subdivisée en plusieurs classes, entre autres celle relative au traitement de la cohérence et la complétude d'une spécification donnée par un utilisateur.

Elle est facilement extensible au fur et à mesure que la méthodologie évolue et est raffinée par l'introduction de règles suggérées par des experts. Le système expert utilisé a pour formalisme de représentation de la connaissance les prédicats et les règles de production. Ce formalisme est connu pour être puissant au niveau de la déduction mais pauvre en sémantique. C'est ce dernier point qui rend difficile la formalisation de la connaissance experte et sa traduction.

Un certain nombre d'extensions est envisagé au système :

- penser à remplasser le formalisme de prédicat par un formalisme tel les réseaux sémantiques,
- augmenter les connaissances qu'il manipule en quantité et en qualité. Par exemple, permettre plus de constructions syntaxiques au niveau de la syntaxe et plus de concepts au niveau sémantique,
- augmenter sa tolérance aux erreurs, et sa flexibilité,
- concevoir d'autres types d'interfaces, par exemple une interface graphique et une interface traitant un langage de spécification qui viendraient s'ajouter à l'interface langage naturel,

- de pouvoir supporter une phase de reverse engineering permettant de passer d'un schéma relationnel à un schéma conceptuel (dans un certain modèle, en particulier l'E/A).

Références :

[ALI 91] Z. Alimazighi, " Evolution des méthodes de développement d'application bases de données ", 1er séminaire sur les bases de données, Alger, Juin 1991.

[AZO 94] A. Azouaou, M.T. Djemai, " Système d'aide à la conception de bases de données en langage naturel ", thèse d'ingéniorat, U. S. T. H. B , 1994.

[BAT 92] C. Batini, S. Ceri, SB. Navathe, " Conceptual database design. An Entity-relationship approach Benjamin/Cummings, 1992.

[BOU 85] M. Bouzeghoub, " Une base de connaissances pour un système expert en conception de bases de données ", actes des journées d'étude (Dijon 1985), collection MBD.

[CAP 80] J. & J. P. Caput, " Dictionnaire des verbes français ", Larousse, 1980.

[CAU 88] C. Cauvet, " Un modèle et un outil d'aide à la conception des systèmes d'information ", Thèse de doctorat, Paris 6, 1988.

[CER 93] " CERESY. Manuel de l'utilisateur", CERIST.1993.

[CHE 92]. M. Chein, M.-L Mugnier, " Conceptual graphs : fundamental notions ", Intelligence Artificielle, Vol. 6, N°. 4, 1992.

[CLE 85] E K Clemons, " Data models and the ANSI/SPARC architecture ". in [YAO 85]

[DIV 92] M Divine, " Parlez-vous Merise ?", 4eme éd., Eyrolles, 1992.

[FAR 89] J. Fargues, " Des graphes pour coder le sens des phrases ", Pour la science, N°. 137, 1989.

[LEV 91] G. Levreau, J.-N. Meunier, M. Bouzeghoub, E. Métais, "Définition d'une interface langage naturel pour la conception de bases de données", Rapport technique MASI, 1991.

[MET 93] E. Métais, J.-N. Meunier, G. Levreau, " Database schéma design, validation and view intégration : a perspective from natural language ", Rapport MASI - Paris VI, 1993.

[MOU 92] B. Moulin, P. Creasy, " Extending the conceptual graph approach for data conceptual modeling ", Data & Knowledge Engineering, Vol. 8, 1992.

[NOU 91] 0. Nouali, "Conception et réalisation d'un système de compréhension de phrases interrogatives et de génération automatique de réponses en langage naturel (SIGAR)", Thèse de Magister, CDTA, CERIST, 1991.

[OBR 88] D. Obretenov, Zh. Angelov, J. Mihaylov, P. Dishlieva, N. Kirova, " A knowledge-based approach to relational database design ", Data & Knowledge engrg., vol.3 (1988), pp. 173-180.

[REI 92] D. Reiner, " Database design tools ", in [BAT 92].

[SAB 89] G. Sabah, "L'intelligence artificielle et le langage. Vol.1, processus de compréhension", 2ème édition,

Hermès. 1989

[SAB 90] G Sabah. "L'intelligence artificielle et le langage Vol.2, représentation des connaissances". Hermès, 1990

[SOW 86] J F. Sowa, E C Way, "Implementing a semantic interpréter using conceptuel graphs ". IBM J Res. Develop. vol 30, n°1, January 1986.

[STO 91] V C Storey, " Relational database design based on the entity-relationship model ", Data & Knowledge engrg.. Vol 7, 1991

[YAO 85] SB Yao (éd.), " Principles of database design, Vol. 1 Logical organizations ", Prentice-Hall, 1985.