1474

E.N.S.S.I.B
ECOLE NATIONALE SUPERIEURE
DES SCIENCES DE L'INFORMATION
ET DES BIBLIOTHEQUES

UNIVERSITE
CLAUDE BERNARD
LYON I

DESS en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage

Adhoc et la Gestion Electronique de Documents : de la Conception à la Commercialisation.

Philippe BREVET

Sous la direction de :

Monsieur **Richard BOUCHE**
E.N.S.S.I.B
Monsieur **Bernard CHARNOMORDIC**
A.I.O
Monsieur **Mohamed HASSOUN**
E.N.S.S.I.B

1992

E.N.S.S.I.B
ECOLE NATIONALE SUPERIEURE
DES SCIENCES DE L'INFORMATION
ET DES BIBLIOTHEQUES

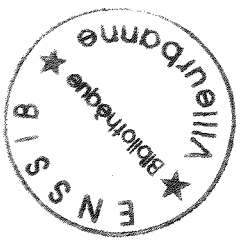
UNIVERSITE
CLAUDE BERNARD
LYON I

DESS en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage

Adhoc et la Gestion Electronique de Documents : de la Conception à la Commercialisation.

Philippe BREVET



Sous la direction de :

Monsieur **Richard BOUCHE**
E.N.S.S.I.B
Monsieur **Bernard CHARNOMORDIC**
A.I.O
Monsieur **Mohamed HASSOUN**
E.N.S.S.I.B

1992
ID
ST 24

1992

Adhoc et la Gestion Electronique de Documents : de la Conception à la Commercialisation.

Philippe BREVET

Résumé : ce rapport concerne un stage du DESS d'Informatique Documentaire, qui s'est déroulé du 1er Mars au 30 Juin 1992 à l'Assistance Informatique de l'Ouest à Dreux (Eure et Loir). Cette SSII développe et commercialise plusieurs logiciels, dont le logiciel documentaire Adhoc. Le stage consistait à seconder l'entreprise dans le développement et le lancement commercial de AdhocPlus, logiciel documentaire fonctionnant en réseau, ainsi que d'assister l'équipe commerciale dans le marché grandissant de la gestion électronique de documents ; je devais également mener une réflexion sur les structures logiques des bases et fichiers documentaires, en particulier sur le modèle relationnel.

Descripteurs : Base Donnée Relationnelle, Documentation Automatique, Information Numérique, Informatique Documentaire, Logiciel.

Abstract : this paper describes a mission (part of the DESS of documentation data processing), which happened between the 1st of March and the 30rd of June of the year 1992. It took place in Dreux (Eure et Loir, France), in the computer software developping society "Assistance Informatique de l'Ouest", which develops and sales a document retrieval software nammed Adhoc. My function consisted in helping in the developpment of AdhocPlus, a new netware based software, and contributing in the sales of electronic based document retrieval systems. An other point was to think of ways to structure information (data bases) for automatic processing, using the relational schemes.

Keywords : Relational Database, Automatic Documentation, Digital Information, Documentation Data Processing, Software.

Je tiens à remercier chaleureusement la sympathique équipe de l'Assistance Informatique de l'Ouest, ainsi que ses dirigeants, sans qui ce stage n'aurait pas eu lieu.

Sommaire

1. L'Assistance Informatique de l'Ouest

Cette entreprise, fondée en 1980 à Dreux, repose sur le développement et la commercialisation de deux types de logiciels :

- un logiciel d'optimisation de découpe, visant le marché des menuisiers, miroitiers, et autres entreprises professionnelles de la découpe. Ce logiciel fut l'un des tous premiers de son genre à faire son apparition en France, et reste aujourd'hui leader sur le marché national. Trois personnes travaillent actuellement à son développement, et deux salariés sont chargés de sa commercialisation, non compté les distributeurs extérieurs.

- un logiciel documentaire, conçu vers 1983, dont les besoins étaient ressentis par plusieurs personnes qui se sont donc associées pour créer Doc'AIO, devenu Adhoc sous l'interface graphique Windows© en 1986, et dont le "grand-frère" AdhocPlus commence à être commercialisé aujourd'hui. Cinq personnes sont chargées de son développement, et la commercialisation est prise en charge par trois salariés.

L'A.I.O a également pour fonction :

- la distribution des produits Apple ;

- la commercialisation des logiciels bureautiques tels que la gamme Microsoft ;

- la vente de matériel micro-informatique de type PC ;

- la distribution de logiciels Novell ;

- le conseil et l'ingénierie informatique.

Son effectif est aujourd'hui d'environ 20 salariés, et son chiffre d'affaire avoisine les 10 millions de francs.

2. Missions du Stage

Deux missions ont en effet dirigées mon stage pendant ces 4 mois :

- la première, qui représente la plus grande quantité de travail, était d'assister l'entreprise dans le développement du logiciel AdhocPlus, ainsi que de me former aux nouvelles techniques de la gestion électronique de documents (GED), en vue de l'expansion du marché de l'informatique documentaire.

Cela consistait donc à : - utiliser les logiciels et les outils de la GED ;

- aider à la conception des différents manuels du logiciel AdhocPlus ;

- tester le logiciel AdhocPlus ("débugger") ;

- mettre au point les notices et les logiciels d'installation Adhoc ;

- pouvoir faire des présentations commerciales des outils de la GED.

- la seconde, beaucoup plus théorique, correspond à long terme à la création d'un ouvrage sur la conception de bases documentaires Adhoc. Cette mission n'a bien sur pas aboutie en quatre mois, mais la lecture de différents documents me permettra, l'an prochain je l'espère, de poser les bases d'un tel ouvrage. La première mission m'a tout de même permi d'appréhender un peu plus le monde des documentalistes, que je ne connaissais que par le biais de l'ENSSIB.

Les questions et les attentes de certains clients d'une part, associées aux possibilités techniques que j'observais lors du développement du logiciel Adhoc (programmation orientée objet) ou de mes lectures techniques (réseaux, gestion d'image, systèmes d'exploitation multi-tâches ...), le tout coiffé par des lectures théoriques sur les bases de données relationnelles, l'architecture client-serveur, me permettront d'aborder prochainement (c'est mon voeux) tout aussi bien les questions pratiques sous un aspect théorique que réciproquement, les questions fondamentales sous un aspect technique.

3. Gestion Documentaire et Gestion Electronique de Documents.

3.1 Les trois phases de la gestion documentaire.

Le service documentaire de l'entreprise peut être schématisé de la sorte :

Hier : un système de gestion de documents.

- une personne trop âgée pour suivre un cours de recyclage, et que l'on avait affecté au service documentaire pour attendre la retraite ;
- des étagères garnies de collections éparses, prises ça et là au grès des dons et des achats, faisant office aussi bien de "bibliothèque publique d'entreprise" que de centre de documentation, les salariés ayant de toute façon accès à l'information les concernant par leurs propres moyens ;
- des classeurs et des fichiers manuels.

Aujourd'hui : un système informatique de gestion de documents.

- une documentaliste performante (mais parfois dépassée par l'imposante quantité d'information) ;
- un logiciel documentaire ou un système de gestion de bases de données, voire une solution "maison", travaillant sur micro-ordinateur en mono-poste, associé à une imprimante matricielle ;
- une photocopieuse, des ciseaux, et de la colle ...

Demain : un système ouvert de gestion électronique de documents .

- une documentaliste toujours performante ;
- un scanner ;
- des logiciels
  - d'OCR (reconnaissance optique de caractères) ;
  - d'indexation de texte intégral ;
  - de gestion d'images ;
  - de PAO (Présentation Assistée par Ordinateur) ;
- un logiciel documentaire ;
- une imprimante laser de type "postscript" ;
- et une configuration réseau local, chaque salarié pouvant accéder aux informations pertinentes à ses yeux (et auxquelles il a le droit d'accéder), informations de type texte, image fixe, voire bientôt séquence sonore, ou film vidéo (ce qui est techniquement faisable aujourd'hui, mais demande trop de ressources matérielles et financières ...).

Ces trois situations, décrites ici d'une façon volontairement caricaturale, ne sont pas fortuites : elles répondent en effet à différentes demandes de la part de l'entreprise, de ses salariés et surtout de ses dirigeants :

1- la première à des petites entreprises, des artisans, ou des organisations plus importantes il y a quelques années, où chacun parvient à obtenir par lui-même le peu d'information qu'il nécessite. Le centre de documentation, s'il existe, tient plus le rôle d'une bibliothèque / salle de lecture

qu'un d'un service stratégique de l'entreprise.

2- les informations que chacun parvenait à obtenir deviennent trop nombreuses, et les salariés n'arrivent plus à les gérer d'une façon satisfaisante. Elles sont donc confiées à un(e) documentaliste spécialisé(e), qui sait donner la bonne référence à l'instant propice.

3- Le volume d'information augmente encore, et la notion de document s'élargit : ce n'est plus seulement quelques feuilles dactylographiées ou un article de presse, mais aussi une image numérisée (photographie, plan, ...), une bande sonore, une séquence vidéo ... Et chacun désire l'obtenir à tout moment, rapidement. L'information documentaire doit donc reposer sur un réseau informatique.

Ces situations correspondent également à différentes situations économiques, avec une concurrence inter-entreprises beaucoup plus forte aujourd'hui, des marchés de plus en plus étendus géographiquement, des outils techniques de plus en plus sophistiqués, et évoluant très rapidement.

Le concept de veille (économique, technologique, ...) est de rigueur aujourd'hui, le sera d'autant plus demain, et le rôle de guetteur revient au personnel des centres de documentation.

3.2 La Gestion Electronique de Documents.

La GED va, dans les années à venir, probablement bouleverser le traitement et le circuit du document dans les organisations, administrations ou entreprises.

Le centre de documentation a pour rôle le choix des abonnements et des ouvrages, la réceptions des documents ; puis est ensuite effectué le dépouillement et l'indexation intellectuelle.

Viennent ensuite les traitements des articles sélectionnés, traitements effectués le plus souvent manuellement. La revue de presse (c'est souvent le media choisi par le centre de documentation pour la diffusion de l'information) résulte des opérations de photocopiage, coupage et collage. C'est ce travail fastidieux qui peut être automatisé en tout ou partie grâce aux outils de la GED.

Un poste de gestion électronique de documents peut comporter :

- un scanner, noir et blanc, couleurs ou nuances de gris ;
- un disque magnéto-optique ;
- un logiciel de gestion d'images, intégrant de préférence le pilotage du numériseur ;
- un logiciel de reconnaissance optique de caractères (OCR), qui peut également commander le scanner ;
- un logiciel d'indexation de texte intégral ;
- un logiciel documentaire.

3.2.1 La numérisation des documents.

Un numériseur, ou scanner, est en quelque sorte un photocopieur reproduisant l'information «lue» , non pas sur papier mais dans un fichier de type «point», le plus souvent au format TIFF.

C'est un outil encore peu répandu, mais on peut trouver aujourd'hui dans le commerce des scanners pour environ 5 000 francs, et des scanners à main pour moins de 2 000 francs. Mais il est certain que, comme dans le cas d'un photocopieur, l'investissement doit être proportionnel à l'usage requis. Un utilisateur numérisant plusieurs dizaines de pages quotidiennement ne pourra se contenter d'un scanner à main, qui peut parfaitement convenir pour une utilisation occasionnelle ou pour des documents d'un format réduit. Pour un usage intensif, une vitesse de numérisation supérieure à 10 pages par minute semble être nécessaire.

Le logiciel de pilotage doit être simple d'utilisation, tout en autorisant des réglages sophistiqués si besoin est. Entre autres, l'utilisateur doit pouvoir paramètrer la définition (nombre de points par pouce, entre 200 et 400 en règle générale), la luminosité, le contraste ... Par exemple, lors d'une numérisation qui sera suivie par une reconnaissance optique de caractères, le papier glacé nécessite un réglage plus sombre qu'un papier ordinaire. Il faut en effet ne pas croire en une solution idéale : bien que les réglages par défaut donnent en général de bons résultats, un apprentissage est nécessaire, comme dans le cas d'un photocopieur.

3.2.2 Le stockage : disques magnéto-optiques.

Cette technologie récente combine les avantages des disques magnétiques (rapidité) et des disques optiques (capacités de stockage importantes). Dans le cas des images numérisées, la mémoire de masse est un facteur important dans la mesure où les fichiers en mode point peuvent rapidement saturer l'espace disque interne du micro ordinateur s'il ne contient «que» 40 ou 80 Mo, configurations classiques dans un PC de type 80386.

Les capacités sont de l'ordre de 400 Mo par face (et augmentent encore), les disques étant double faces. Ces disques sont facilement amovibles, et d'un coût raisonnable.

Le disque magnéto-optique n'est pas actuellement destiné à remplacer le disque magnétique (trop lent) mais peut être configuré comme mémoire de masse secondaire, dans un réseau local, comme volume d'un serveur image par exemple.

3.2.3 Le logiciel de gestion d'images.

Celui-ci doit de préférences intégrer le pilotage du numériseur pour éviter des passages entre logiciel. En effet, l'image numérisée occupe beaucoup d'espace mémoire : environ 1 Mo pour une page A4 avec une définition de 300 DPI (Dot Per Inch) en noir et blanc, et 2 Mo en 400 DPI. On ne peut donc ouvrir plusieurs applications en manipulant de telles quantités d'informations sans disposer d'une mémoire RAM importante, voir imposante.

Il est également souhaitable que le logiciel puisse compresser les fichiers de données, de préférence à l'aide d'algorithmes normalisés (groupes III et IV des télécopieurs par exemple). Répondant à des normes, ces fichiers pourront ainsi être reconnus par d'autres logiciels (e.g., le logiciel OCR OmniPage ne reconnait pas les fichiers compréssés par le logiciel de pilotage des scanners Hewlett Packard).

La gestion d'image est également nécessaire, et des fonctions d'indexations peuvent être très utiles car le nom de fichier, en particulier sous DOS (limité à 8 caractères), ne permettra pas de nommer explicitement un document, et l'utilisateur sera vite désorienté. Des zones d'édition de textes doivent donc pouvoir être liées à l'image pour une description par exemple, et une recherche sur mots-clefs doit être possible. Un classement hiérarchique des fichiers de type Armoire/Tiroir/Dossier peut être très utile en cas de quantités importantes de documents numérisés.

La gestion graphique de l'image peut elle-aussi être une fonction intéressante, permettant de ne conserver que la zone pertinente de l'image (que ce soit un texte ou une photographie). Les fonctions standard issues du Macintosh telles que couper-copier-coller, que l'on retrouve dans l'interface Windows®, permettent également une économie de mémoire de masse. Les articles de périodiques sont souvent entrecoupés d'encarts publicitaires.

3.2.4 Le logiciel de reconnaissance optique de caractères.

La reconnaissance optique de caractères est doublement intéressante dans système d'information documentaire :

- le texte reconnu peut être modifié grâce à un simple traitement de texte, ou être traité par un logiciel d'indexation intégrale, ... Le texte «OCéRisé» acquiert ainsi une valeur «sémantique».
- une page de texte (sous forme ASCII ou autres) ne dépasse pas quelques Kilo-octets, d'où un gain important de place en mémoire de masse.

Le logiciel OCR doit parvenir à un taux de reconnaissance proche des 100% (supérieure à 95%) pour avoir une utilité dans la chaîne documentaire. Les différentes polices de caractères classiques doivent être reconnues, et l'utilisateur doit pouvoir créer de nouveaux fichiers de polices (symboles mathématiques, chimiques ...). Le format (gras, italiques ...) et la taille sont également des paramètres intéressants à conserver.

Le logiciel d'OCR doit de préférence être en mesure de piloter des numériseurs, de sélectionner une partie de la page à numériser pour ainsi économiser de l'espace disque en ne conservant que l'article, sans tenir compte des encarts publicitaires et autres formes de bruit.

Des fonctionnalités telles que la possibilité de voir l'image numérisée au sein du logiciel d'OCR sont très utiles pour la correction des erreurs ; on peut ainsi voir le mot dans son contexte. Il est également intéressant de pouvoir conserver les graphiques, les logos, les photos, ... en mode point dans un fichier de format TIFF, pour pouvoir ainsi le réutiliser dans un logiciel de type pageur (PageMaker, Xpress, ...) ou traitement de texte perfectionné.

3.2.5 Le logiciel d'indexation intégrale de textes.

Le texte issu du traitement de l'OCR, ainsi que des fichiers provenant de traitements de texte (courrier, documentation technique, ...), pour être exploités efficacement, doivent être indexés. L'indexation intellectuelle n'est pas toujours possible (manque de temps, de personnel qualifié, ...), et n'est pas toujours de rigueur : par exemple, le milieu juridique dispose d'un langage fortement normalisé, des mentions précises (numéro d'article d'un code civil, pénal, ...), et une indexation en texte intégral sur de nombreux documents peut permettre un gain appréciable de temps, ainsi qu'offrir de nombreuses possibilités de recherche.

Les fonctions de recherche doivent être développées de manière à permettre l'utilisation d'opérateurs booléens, les opérateurs de comparaison et de proximité. Les stratégies de recherches doivent pouvoir être mémorisées pour permettre la diffusion sélective de l'information dans l'entreprise.

Le logiciel d'indexation intégrale doit être en mesure de gérer des documents de plusieurs Méga-octets, tel que des conférences, des ouvrages, ... Et l'index doit pouvoir contenir de nombreux documents (fichiers), créant ainsi une base documentaire en texte intégral.

Un point important est la non-duplication des documents existants : si le fichier initial est dans un format Word, l'action du logiciel d'indexation intégrale ne doit en aucun cas l'altérer, ni le dupliquer pour un évident gain de place sur la mémoire de masse. Mais l'utilisateur doit pouvoir accéder au document de type Word à partir du logiciel d'indexation. Il est donc important qu'il supporte de nombreux formats de traitement de texte, qui ne sont pas encore normalisés, bien qu'une lueur d'espoir apparaisse avec la généralisation du format RTF (Rich Text Format).

L'indexation "quantitative" n'est pas destinée à remplacer l'indexation "qualitative", mais elle peut favorablement la compléter ; le travail intellectuel du documentaliste se poursuit par l'utilisation d'un logiciel documentaire.

3.2.6 Le logiciel documentaire.

Le logiciel documentaire est l'outil fédérateur des applications précédemment citées. Le système de fiches en cartons sur lesquelles sont inscrites les rubriques est alors remplacé par un écran informatique. Mais ceci ne justifierait en rien sa nécessité, au contraire.

L'outil informatique apporte, chose essentielle dans le cadre de la gestion documentaire, des fonctions d'interrogation puissantes, n'entraînant pas comme dans le cas du traitement manuel, une multiplication des fichiers. En effet, dans le cas d'un fichier manuel, trois possibilités d'accès (titre, auteurs, matières) impliquent trois fichiers physiques. Dans le cas des fichiers informatiques, ces champs seront probablement soumis à une indexation, et il y aura donc un léger accroissement du volume. De plus, toutes les procédures de tri seront automatisées.

Les fonctions d'interrogation doivent utiliser les opérateurs booléens, les opérateurs de comparaison, les troncatures et masques, et l'indexation en texte intégral peut également apporter un bénéfice lors de recherches sur des champs contenant des textes, tels que le titre ou le résumé. Des recherches combinées sur plusieurs fichiers peuvent être nécessaires dans certains cas de figure, voir des recherches multi-bases. Le logiciel doit également permettre de sauvegarder les interrogations pour favoriser la diffusion sélective de l'information.

Un facteur important de l'interrogation est la rapidité de recherche, qui est souvent fonction de l'organisation de la base. Le logiciel doit donc posséder des fonctions d'administration de base puissantes, tout en restant conviviales, permettre les liens inter et intra-fichiers, l'indexation intégrale. Une possibilité d'évolution de la base sans perte de données est nécessaire. La notion de réciprocité des liens existant dans le logiciel Adhoc est intéressante pour la constitution de fichiers de type thesaurus, et évite de nombreuses saisies.

La saisie, phase importante de la gestion documentaire, peut être agrémentée par des insertions automatiques, des choix dans les listes d'autorité, mais elle reste toutefois la tâche la plus fastidieuse. L'interface Windows©, pour les compatibles PC, ou l'interface graphique du Macintosh, apportent les fonctions de couper-copier-coller qui peuvent éviter la ressaisie d'un texte existant. Par exemple, les résumés d'ouvrages situés en dos de couverture peuvent parfaitement être numérisés, OCéRisés, puis ensuite copiés dans un champ du fichier documentaire.

Il est souhaitable également dans le milieu documentaire de disposer de fonctions d'impressions élaborées. Le logiciel Adhoc est maintenant livré avec le "générateur de modèles d'impression", qui permet à l'utilisateur des impressions sophistiquées (champs de plusieurs

fichiers au sein d'une même notice, combinaison de différentes polices, différents formats, différents styles, positionnement des champs dans la page pour obtenir une réprésentation graphique d'une hiérarchie, dans le cas d'un thesaurus, ...).

Le logiciel documentaire doit aujourd'hui offrir des services complémentaires, liés au multimedia. Deux solutions sont possibles : le logiciel intègre des fonctions de gestion d'image, de pilotage de numériseur, ... , et offre ainsi une solution complète. Autre politique possible, le logiciel utilise les services de logiciels spécialisés, grâce aux nouvelles formes de programmation sous Windows® : utilisation des Dynamic Data Exchange (DDE), Dynamic Link Library (DLL), et, dernière née, l'OLE (Object Linking and Embedding) qui permet de créer des documents composites, ce qui signifie que la modification d'un tableau dans Excel répercutera automatiquement le changement dans le logiciel (traitement de texte par exemple) qui utilise l'objet.

C'est ce choix qui a été effectué par l'AIO dans la conception de AdhocPlus, laissant aux différents fournisseurs de logiciels (de gestion d'images entre autres) la tâche qu'ils maîtrisent parfaitement, et qu'ils sauront faire évoluer, les deux applicatifs étant reliés à postériori, et seulement si nécessaire, par le biais des liens dynamiques.

C'est cet aspect fédérateur plus qu'intégrateur qui évitera la construction d'une "usine à gaz", qu'une petite structure telle que l'AIO ne saura faire évoluer, elle en est consciente, au grès des évolutions technologiques.

3.3 Le réseau local d'entreprise.

Un réseau local est un ensemble de logiciels et de matériels reliés entre eux pour permettre ainsi un partage des ressources matérielles (imprimantes, mémoire des masse, ...), logicielles, et surtout de l'information (bases de données) dans l'organisme, ce qui tend ainsi à améliorer la productivité de chacun.

Il peut s'agir de deux micro-ordinateurs reliés à une imprimante, ou de plusieurs centaines de postes connectés à un gros système.

Le passage de configurations mono-postes à une configuration réseau ne va pas sans difficultés pour l'organisme. Bien que cette transition soit souvent transparente pour l'utilisateur, celui-ci se voit tout de même imposer certaines règles par l'administrateur du réseau, et les mots de passe deviennent nécessaires. La sécurité des différentes bases doit être assurée, les fichiers administratifs entre autres, n'étant pas accessibles à tous.

L'émergence de protocoles de communication (IPX, TCP/IP, ...) adoptés, non pas encore à l'unanimité, mais par un grand nombre de constructeurs, permet de concevoir des réseaux alliant des matériels et logiciels (e.g., systèmes d'exploitation) hétérogènes. Le matériel que l'organisme possède n'est donc pas remplacé automatiquement, ce qui nécessiterait des augmentations budgétaires. Il est possible aujourd'hui de connecter des Macintosh sur un réseau Ethernet de PC, de faire communiquer un réseau Local Talk et un réseau Token Ring par l'intermédiaire d'un routeur, ...

Le gestionnaire de réseaux Netware de la société Novell (environ 70% du marché mondial des réseaux locaux) offre ces différents services, ainsi que d'autres tels que la messagerie, le pilotage d'un onduleur, les queues d'impressions, ...

Il est possible aujourd'hui de se lancer dans une "politique réseaux" sans trop de risque, en choisissant des solutions évolutives. Il est préférable de débuter par un service de l'entreprise, pour ainsi tester la solution technique d'une part, mais également l'acceptation et l'intérêt des utilisateurs, puis évoluer graduellement vers une solution globale.

Dans le cas d'un service documentaire, le réseau est "le moyen d'expression" idéal. L'information peut ainsi devenir accessible à tous et à tout moment.

Une seule station de GED est suffisante dans un grand nombre de cas, celle-ci étant reliée à un serveur de l'entreprise. Il fort probable que l'investissement lié au réseau doive être amorti par d'autres services de l'organisme ; le centre documentaire est trop rarement considéré comme un service critique de l'entreprise.

L'administrateur du logiciel documentaire doit alors procéder à une étude des besoins en informations des différents services de l'entreprise, pour être en mesure de prédéfinir des requêtes : en effet, tous les salariés ne peuvent être formés à l'utilisation du logiciel documentaire. Bien que les interfaces graphiques rendent souvent l'apprentissage plus aisé, l'interrogation reste un domaine où la documentaliste excèle, et des requêtes complexes, utilisant opérateurs booléens, troncatures et parenthésage, ne sont pas à la porté de tous les salariés. Des recherches qui

n'aboutissent pas, en raison d'un manque de compétence du salarié (qui accusera pourtant le logiciel, la machine, voire la documentaliste ...), risquent fort de faire échouer l'intégration du service documentaire.

La conception de ce système d'information doit de préférence reposer sur une étude de marketing documentaire. En effet, le besoin des salariés existe-t'il réellement ? Est-ce une volonté de la Direction ? Plusieurs paramètres sont donc à prendre en compte. L'informatique est un moyen et non une fin pour l'entreprise utilisatrice, or certains directeurs informatiques ne partagent pas cette vision, et peuvent engager l'organisme dans un projet dont l'utilité n'est pas fondée.

4. Prémices à la Conception d'un Futur Logiciel Documentaire ?

4.1 Les systèmes de gestion de bases de données relationnelles (SGBDR).

Un SGBD est essentiellement caractérisé par le modèle de données qu'il supporte. Deux générations coexistent aujourd'hui :

- la première, qui correspond aux modèles hiérachique et réseau, et
- la seconde, basée sur le modèle relationnel.

Le **modèle relationnel**, qui nous intéresse ici, a pour fondement les théories mathématiques du même nom, mises en avant par Codd, chercheur chez IBM, au début des années 1970, les systèmes commerciaux étant disponibles depuis 1980.

Deux objectifs sont à l'origine du modèle :

- «Permettre un haut degré d'indépendance des programmes d'applications et des activités interactives à la représentation interne des données, en particulier au choix des ordres d'implantation des données dans les fichiers, des index et plus généralement des chemins d'accès.

- Fournir une base solide pour traiter les problèmes de cohérence et redondance des données.

Ces deux objectifs qui n'étaient pas atteints par les modèles réseau et hiérarchique, ont été pleinement satisfaits par le modèle relationnel, d'une part grâce à la simplicité des vues relationnelles qui permettent de percevoir des données sous forme de table à deux dimensions, et d'autre part grâce à la théorie de la normalisation.» [Gardarin89]

(Pour plus de détails sur la normalisation voir [Koutchouk89] et [Delobel82]).

Le modèle relationnel a également permi le développement de langages de manipulation de données ensemblistes basés sur des théories solides (grâce à l'algèbre relationnel et des langages assertionnels, tel que SQL).

Les structures de données du modèle relationnel reposent sur trois notions de base :

- **le domaine** : ensemble de valeurs caractérisé par un nom ;

- **la relation** : sous-ensemble du produit cartésien d'une liste de domaine caractérisé par un nom ;

- **l'attribut** : colonne d'une relation caractérisée par un nom.

La représentation la plus fréquente du modèle relationnel est une table à deux dimensions.

Six opérations de bases peuvent être effectuées sur ces relations :

- **l'union** : opération portant sur deux relations de même schéma RELATION1 et RELATION2 consistant à construire une relation de même schéma RELATION3 ayant pour tuples ceux appartenant à RELATION1 ou RELATION2 ou aux deux relations ;

- **la différence** : opération portant sur deux relations de même schéma RELATION1 et RELATION2 consistant à construire une relation de même schéma RELATION3 ayant pour tuples ceux appartenant à RELATION1 et n'appartenant pas à RELATION2 ;

- **le produit cartésien** : opération portant sur deux relations RELATION1 et RELATION2 consistant à construire une relation RELATION3 ayant pour schéma la juxtaposition de ceux des relations opérandes et pour tuples toutes les combinaisons des tuples des relations opérandes ;

- **la projection** : opération sur une relation RELATION1 consistant à composer une relation RELATION2 en enlevant à la relation initiale tous les attributs non mentionnés en opérande (aussi bien au niveau du schéma que des tuples) et en éliminant les tuples en double qui sont conservés une seule fois ;

- **la restriction** : opération sur une relation RELATION1 produisant une relation RELATION2 de même schéma mais comportant les seuls tuples qui vérifient la condition précisée en opérande ;

- **la jointure** : opération consistant à rapprocher selon une condition les tuples de deux relations RELATION1 et RELATION2 afin de former une troisième relation RELATION3 qui contient l'ensemble de tous les tuples obtenus en concaténant un tuple de RELATION1 et un tuple de RELATION2 vérifiant la condition de rapprochement. La **jointure naturelle** élimine la redondance de tuples.

(Toutes ces définitions sont issues de [Gardarin89]).

Sept opérations complémentaires peuvent être définies :

- **l'intersection** ;

- **la division** ;

- **le complément** ;

- **l'éclatement** ;

- **la jointure externe** ;

- **la semi-jointure** ;

- **la fermeture transitive**.

(Voir [Gardarin89], [Flory87] ou [Delobel82] pour un approfondissement de ces notions).

4.2 Langages assertionnels et architectures clients / serveur.

Ces langages «permettent de définir de manière non procédurale des suites d'opérations de l'algèbre relationnel» [Gardarin89].

Ils sont directement utilisables par les usagers, ne nécessitant pas de structuration particulière (car non procéduraux). Les langages les plus célèbres sont :

- SQL : Structured Query Language ;

- QUEL ;

- QBE : Query By Exemple.

Ils permettent tous trois les quatre opérations élémentaires effectuées sur les bases de données :

- la recherche ;

- l'insertion ;

- la suppression ;

- la modification.

De plus, ces langages sont bien adaptés pour gérer les **contraintes d'intégrité**. Des «**triggers**» peuvent être attribués à des tables. Lors d'un accès à cette table pour y effectuer une opération (quelle qu'elle soit) le «trigger» entre automatiquement en action pour faire respecter ces contraintes définies auparavant. Il est donc aisé de constituer un réseau basé sur une architecture clients-serveur, en associant également des **procédures stockées** sur le serveur.

En effet, «avec un serveur de réseau classique, lorsque l'un des utilisateurs cherche une information contenue dans la base de données, il charge sur son poste l'ensemble de la base, quelque soit sa taille. Puis le poste assure la recherche.» [Desmedt90]

Ce mode de fonctionnement implique donc :

- un gros traffic sur le réseau, et
- une sécurité des données plus ou moins assurée.

L'architecture **client-serveur** permet donc de réduire le traffic réseau au minimum, et l'intégrité de la base de données est assurée. Le poste client, qui gère l'interface du logiciel, envoit au serveur les paramètres de l'action à réaliser (e.g., critères de recherches), qui traite la demande grâce aux procédures stockées, puis retourne la réponse.

On peut associer à cette architecture le mode transactionnel, pour assurer ainsi une plus

grande sécurité. «Une transaction est une unité logique de traitement qui consiste en une ou plusieures instructions SQL considérées par le gestionnaire de la base de données comme une entité unique. Toutes les instructions constituantes seront éxécutées, ou bien aucune ne le sera» [Pasleau89].

En cas d'incident (e.g., coupure de courant) lors de la transaction, celle-ci n'aura pas modifié l'état de la base, ce qui pourrait avoir des conséquences facheuses.

Par exemple, si un client (dans le cas d'une banque) veut effectuer un virement, deux opérations sont menées :

- débit d'une somme S du compte A ;
- puis crédit d'une somme S du compte B.

Or si l'incident a entrainé l'absence de crédit, la somme S sera "perdue".

Dans le cas du transactionnel, le système conserve toutes les transactions dans un journal, et lors de la reprise, il effectuera automatiquement la restauration de la base («RollBack»).

De plus, le serveur ne verrouillera l'accès qu'à l'unité supérieure à la demande du client (page, table, ...), d'ou une possibilité d'accès par les autres utilisateurs aux données non liées à la transaction. Il existe également la possibilité de définir des bases de données réparties sur plusieurs serveurs.

(Pour un bref aperçu du transactionnel, voir [Poussinov90]).

4.3 SGBD et recherches actuelles.

Les bases de données réparties sont avec les bases de données objets et déductives ce que l'on entend par troisième génération.

- Les **bases de données objets** : «une application a un aspect statique, représenté par les données, et un aspect dynamique, représenté par les traitements sur ces données. (...) les bases de données orientées objets apportent des avantages qui pourraient être déterminants : elles permettent de prendre en compte la dynamique associée aux objets sous forme d'opérations stockées avec les objets et apportent par là une réponse partielle au problème de réutilisation du code des programmes.»

Celles-ci peuvent s'avérer indispensables pour le stockage et la manipulation des objets complexes tels que les textes, les programmes, ...

- Les **bases de données déductives** : «Alors que les faits stockés dans les relations constituent la base de données extensionnelle, les règles exprimant les connaissances sous forme de prédicats dérivés (ou relations déduites) constitueront la base de données intentionnelle. L'objectif est d'intégrer un maximum de connaissances dans la base intentionnelle, de manière à réduire le volume de programmes spécifiques à la mise en oeuvre de chaque application et à faciliter le partage rationnel des connaissances.»

- Les **bases de données réparties** : «Une base de données répartie est une collection de données logiquement corrélées et physiquement réparties sur plusieures machines interconnectées par un réseau de communication. Un programme d'application qui manipule une base de données répartie peut alors accéder des données résidant sur plusieures machines, sans que le programmeur ait à connaître la localisation de ces données.»

(Les définitions ci-dessus proviennent de [Gardarin90], ouvrage à consulter pour compléments).

Les recherches sur cette troisième génération ont débuté au début des années 80, et plusieurs systèmes commerciaux sont aujourd'hui disponibles.

4.4 Les bases de données documentaires.

De nombreux ouvrages traitent de la conception des bases de données, mais aucun à ma connaissance ne traite du cas de l'information documentaire. La gestion de celle-ci me semble pourtant différer de la comptabilité, des stocks, ... En effet, un centre de documentation a pour objet d'intégrer de l'information dans une base de données, mais surtout d'être en mesure de la retrouver par une recherche.

Alors qu'un produit se caractérise le plus souvent par un numéro bien défini, il serait peu pertinent de faire rechercher un ouvrage dont on ignore l'existence par un numéro ISBN ! (la base Electre sur Minitel, a d'ailleurs relégué le numéro ISBN de la première à la cinquième position lors de la redéfinition de l'interface). La base documentaire doit donc essentiellement être organisée en fonction des recherches qui seront effectuées.

Dans le cas des bases de données relationnelles, certaines directives sont à suivre, telle que la normalisation des tables. Or ces concepts posés par Codd sont issus en droite ligne des théories mathématiques ensemblistes, qui ne sont pas généralement intégrées dans la formation des documentalistes. On le ressent à la lecture d'un quelconque ouvrage sur le sujet.

La méthode MERISE, utilisée pour déterminer un modèle entité-association, n'est pas non plus au chevet de beaucoup de documentaliste à ma connaissance.

Les différentes phases conceptuelles pour constituer une base de données peuvent être résumées de la façon suivante :

- procéder à l'étude de l'existant, en insistant sur le schéma des flux entre les différents services. Cette étude peut bénéfiquement être réalisée par une personne extérieure (audit, stagiaire, ...) à l'entreprise, dont la vision de l'organisme sera moins subjective.

- puis vient la phase de la conception des entités, à partir de l'étude de l'existant (mais aussi grâce à des rencontres avec le personnel, qui peut remettre profondément en cause cette étude, en particulier si elle a été réalisée par un salarié). Ces entités devraient être normalisées par le concepteur du système d'information pour être volontairement dénormalisées plus tard si le besoin s'en fait sentir (simplification des traitements). De plus, tous les systèmes relationnels commerciaux ne sont pas aptes à offrir des vues multi-tables.
Les entités sont issues du regroupement des unités élémentaires d'informations, répertoriées dans le dictionnaire des données. L'entité, ou objet, est caractérisée par un nom et un ensemble de propriétés. La relation possède également un nom mais peut ne pas contenir de propriétés.
L'objet possède un identifiant unique. Pour la relation, ce dernier est le produit cartésien des identifiants des différentes entités qu'elle associe.
Il s'agit également de déterminer les cardinalités des différentes entités.

- la définition conceptuelle des traitements, qui nécessite l'analyse des procédures existantes et leur modélisation. Elle a pour but la description en termes abstraits mais fidèles d'une certaine réalité des processus de gestion d'une organisation.

- les procédures organisationnelles : il s'agit alors d'associer les traitements aux utilisateurs, en prenant en compte la notion temporelle et la notion de lieu.

(Pour une approche de la démarche de conception d'un SI, voir [GALACSI86] et [Flory87]).

4.4.1 Adhoc d'hier à demain.

Le logiciel Adhoc est né des besoins d'un groupe associant un conservateur de la bibliothèque nationale et plusieurs chercheurs, qui nécessitaient pour leur usage professionnel un logiciel documentaire d'utilisation facile sous Windows©. Adhoc a été développé entre 1984 et 1987, année de sa première commercialisation.

C'est un logiciel qui offre de nombreuses possibilités dans la définition des structures de bases documentaires, telles que :

- les champs répétitifs ;

- les multi-champs ;

- les liaisons inter et intra-fichiers ;

- les liaisons réciproques.

On peut donc définir des fichiers de type thesaurus, des listes d'autorité, ...

Mais Adhoc repose sur une base de données "maison" hors des grands standards, et n'a aucune possibilité de gestion d'image, et ne répond donc pas à la demande du multi-média. Pour les mêmes raisons, Adhoc ne peut travailler réellement en réseau (plusieurs postes de saisie).

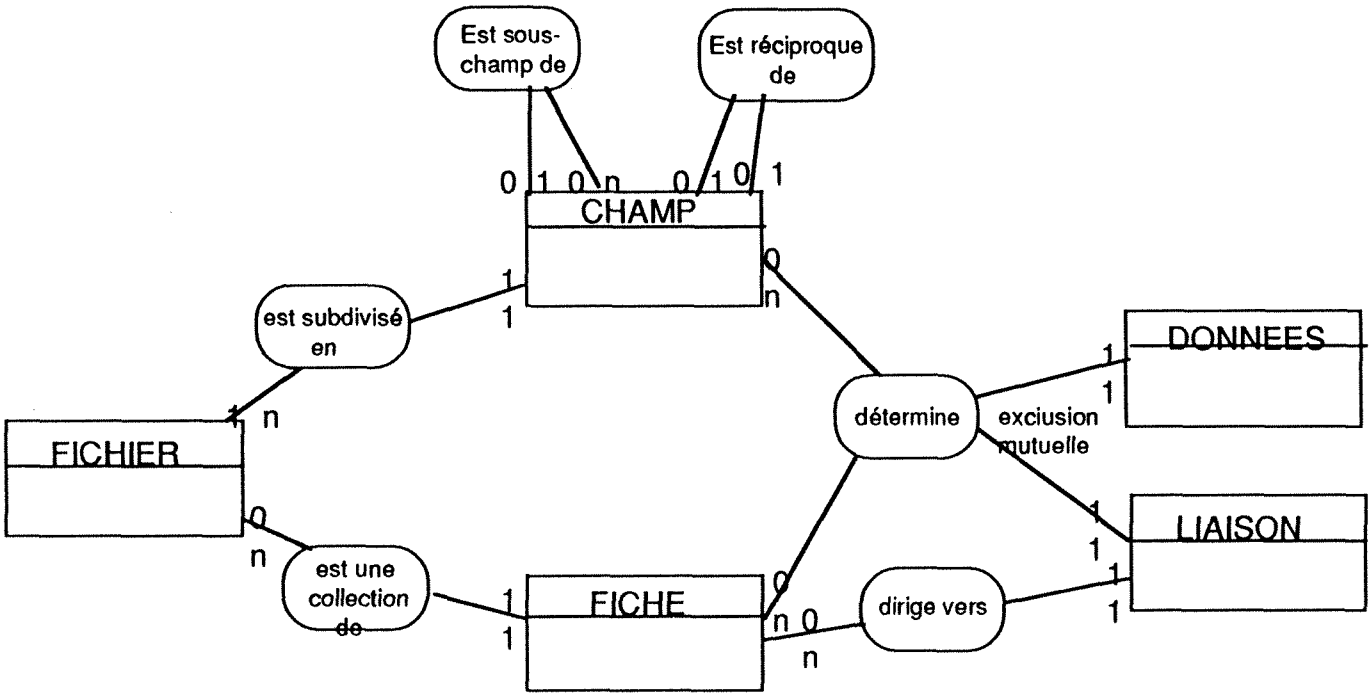
C'est essentiellement pour cela qu'ont été développés AdhocPlus et AdhocSQL, depuis 1990, et dont la commercialisation a débuté en 1992.

Ces deux produits reposent respectivement sur les bases de données DB-Vista de Reima, et SQL Server de Microsoft, et sont donc pleinement compatibles réseaux. De plus, l'interface a été conçue pour utiliser le potentiel de Windows©, et les deux nouveaux produits gèrent les liens externes (possibilité d'appeler un document et son applicatif à partir d'un champ Adhoc, document de type image, dessin, traitement de texte, ...) et répond donc pleinement aux besoins du multi-média.

Bien que AdhocPlus et Adhoc SQL offrent de belles perspectives, certaines limites des produits semblent être apparues dans certains cas spécifiques :

- la recherche dans de grosses bases de données (plusieurs centaines de milliers de fiches documentaires) est relativement longue ;

- la consultation des tables SQL de la base documentaire (créée à partir d'AdhocSQL) semble difficile, et la modification presque impossible si l'on désire utiliser un outil autre que AdhocSQL (tableur par exemple). Il est également difficile d'interroger directement la base de données en SQL sans faire appel à l'interface Adhoc.



MCD simplifié des bases Adhoc

Réalisé par J.P. Lierville

Ceci est partiellement dû au fait qu'une base documentaire Adhoc est représentée dans la base de données par une méta-structure (voir MCD simplifié d'une base Adhoc et la description des tables SQL en annexe).

Une version future (que l'on qualifiera de directe) pourrait donc générer une structure SQL proche du modèle conceptuel des données, ce qui simplifierait les traitements tels que la recherche, et permettrait une consultation / modification des données à partir d'un outil différent de Adhoc (tableur ou directement en SQL).

En effet, une requête sur un champ nécessite dans le modèle actuel la lecture d'un minimum de trois tables, alors qu'un dans le modèle direct, la lecture d'une seule table peut être suffisante (voir MCD "direct" de la base PALAIS).

En contrepartie, la création de base nécessite une démarche d'expertise qui doit donc être formalisée. La démarche de l'interface génératrice pourrait ressembler à celle d'un système expert, en "demandant" à l'administrateur de la base certaines précisions sur les entités et les relations de la base documentaire.

4.4.2 Typologie des champs Adhoc.

Un champ Adhoc peut être :

- un multi-champs : champ composite, tel qu'une adresse (rue, code postal, ville, ...) ;

- un lien vers un fichier Adhoc : lien entre un champ auteur d'un fichier Articles et un fichier des auteurs ;

- un lien externe : lien vers un fichier traitement de texte, ou un fichier image, et qui appelle également l'applicatif générateur ;

Ces champs peuvent avoir comme caractéristique :

- la répétitivité ;

- dans le cas d'un lien, la réciprocité.

4.4.3 Structure Adhoc et Modèle Entités-Association.

Prenons l'exemple d'un besoin documentaire "classique" (cette structure proposée n'est pas fortuite, il s'agit de la base de démonstration livrée avec les logiciels de la gamme Adhoc).

Un service documentaire dépouille quotidiennement des revues périodiques et crée donc pour chaque article sélectionné un ensemble de fiches (notices) :

- une fiche Article, contenant un numéro d'inventaire, le titre, la date de saisie, la source, le ou les auteurs, et les mots-clefs, ... , caractérisant l'article ;
- une fiche Auteur, qui contient des renseignements tels que les prénoms, la nationalité, la date de naissance, ...
- une fiche d'un fichier Thesaurus, qui a pour terme l'un des mots-clefs caractérisant l'article, mais également son ou ses génériques (si thesaurus polyhiérarchique), ses spécifiques, les synonymes, les termes associés, ...

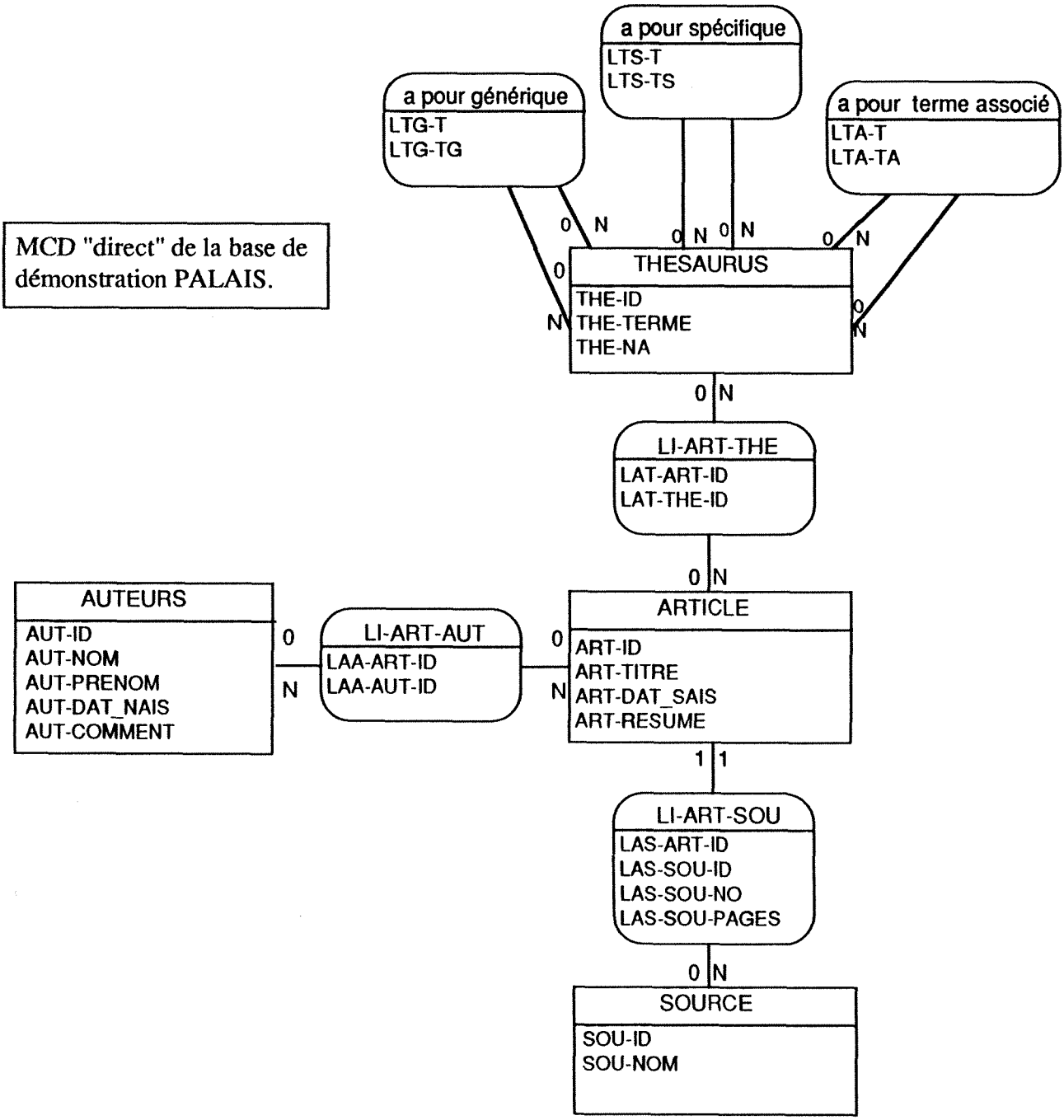
La base peut également contenir un fichier des sources, qui contient les caractéristiques des revues que le centre documentaire reçoit périodiquement.

(Acceptons cette structure telle qu'elle se présente. On aurait pu en effet créer une seule fiche, en ne nous intéressant pas aux auteurs, et ne disposant pas d'un thesaurus.)

Examinons maintenant les différents types de champs.

Un champ répétitif Adhoc (auteur ou mot-clef dans le fichier Article) est, comme son nom l'indique, un champ qui peut avoir plusieurs occurrences. Mais ce nombre n'est pas connu à l'avance par l'administrateur de la base de données. Un champ de type AUTEUR peut contenir aucun auteur (ouvrage anonyme) ou N auteurs. Il est possible de fixer un nombre maximum, en

MCD "direct" de la base de démonstration PALAIS.



considérant par exemple qu'un article peut être écrit par trois personnes au plus, et que les suivantes n'ont pas d'importance.

Ceci peut être considéré comme une règle de gestion de la base de données, qui s'appliquera à 99% des articles saisis, et il suffirait alors de définir trois champs dans le fichier des articles (AUTEUR1, AUTEUR2, et AUTEUR3). Mais cette règle ne reflète pas la réalité.

Les logiciels Adhoc permettent donc d'entrer un nombre "infini" d'occurrence.

Mais qu'est-ce qu'un champ répétitif Adhoc dans le modèle entité-association ?

Il s'agit d'un ensemble entité-association-entité de cardinalité N-M ou N-1.

4.4.4 Modèle Entités-Association et Modèle Relationnel (tables SQL).

Trois règles régissent le passage du MCD aux tables SQL en fonction des occurrence :

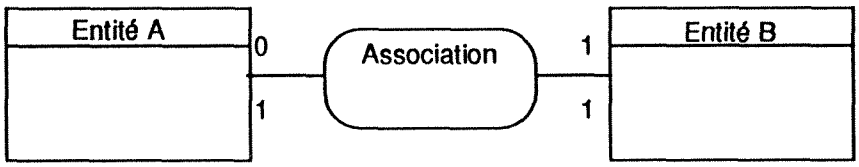
- s'il s'agit d'une association de type 1-1, une seule table SQL est créée ;

- s'il s'agit d'une association de type 1-N, la clef de B et les attributs de l'association sont importés dans A ; deux tables SQL représentent donc le MCD.

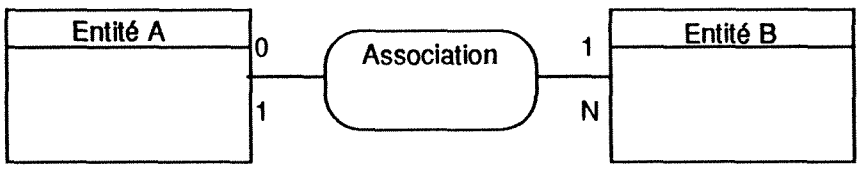
- s'il s'agit d'une association de type N-M, trois tables SQL sont créées.

Une question vient alors : en cas de choix d'une association de type 1-N, faut-il créer trois tables SQL (cas N-M) en prévision d'un changement dans la structure (un champ Adhoc devenant répétitif par exemple) ?

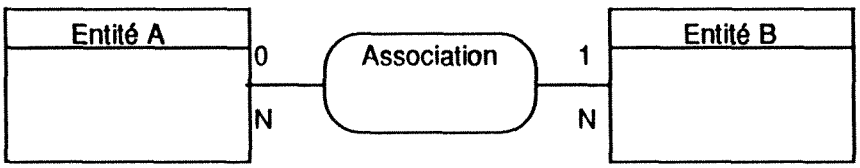
En effet, considérons l'entité A comme étant Article et B comme Auteur, l'association pouvant être «écrit / est écrit ».



-Dans ce cas, qui ne sera pas pris en considération, l'auteur ne peut écrire qu'un seul article, un article étant écrit par un auteur, ou anonyme. C'est une association de type 1-1.



- Dans ce premier cas un article est écrit par 0 ou 1 auteur. Un auteur peut avoir écrit de 1 à N articles. C'est une association de type 1-N.



- Dans ce second cas, un article peut être écrit par plusieurs auteurs. C'est une association de type N-M.

Il s'agit donc bien, dans la structure Adhoc d'une modification des caractéristiques du fichier Article, le champ auteur devenant répétitif.

Le **multi-champs Adhoc** est un ensemble de sous-champs qui forme un sous-ensemble cohérent au sein d'un fichier. Dans notre exemple précédemment cité, la source du fichier article est un multi-champs. L'information de type adresse (numéro, rue, code postal, ville, ...) est généralement contenue dans un multi-champs. Adhoc présente un multi-champs décalé vers la droite par rapport à un champ, ce qui le rend rapidement identifiable.

Un multi-champs Adhoc, dans le modèle entité-association, peut être représenté dans le MCD par une association de type **1-1**, qui ne formera donc qu'une seule table SQL si l'on suit la règle précédemment citée. Si ce **multi-champs** est **répétitif** (e.g. une fiche d'un fichier des personnes peut avoir plusieurs adresses), l'association sera alors de type **N-1**. En effet, sauf exception, une adresse ne correspond qu'à une seule personne. (Il est à noter qu'un multi-champs ne peut être répétitif dans le logiciel Adhoc Standard).

La structure sera dans ce cas représentée en SQL par deux tables, la clef de l'entité Personnes étant importée dans l'entité Adresses. Ici encore se pose la question de savoir s'il faut "anticiper" sur un choix futur de l'administrateur de la base en créant (au niveau de la structure SQL) deux tables alors que l'association est de cardinalité 1-1.

Il est donc important de fixer ce type de règle pour élaborer un générateur de structures SQL (que l'on appellera le Module d'Administration Adhoc).

En résumé :

- pour une association de type 1-N, Adhoc nécessite deux fichiers, le premier ayant un champ lié au second ;

- pour une association de type N-1, créer un multi-champs répétitif dans un fichier Adhoc (AdhocPlus ou AdhocSQL, Adhoc Standard n'offrant pas cette possibilité) ;

- pour représenter une association de type N-M, il faut créer deux fichiers Adhoc, un champ du premier fichier étant répétitif et en liaison avec le second fichier ;

On voit donc qu'une liaison de type 1-N n'est pas la symétrique d'une liaison de type N-1. Il existe une orientation nécessaire dans la conception des fichiers Adhoc. Les fichiers doivent être donc être hiérarchisés en fonction des différents traitements qui seront effectués sur la base.

Certains traitements sont également important. Le logiciel de gestion de bases de données documentaires ne repose pas uniquement sur la constitution d'une structure cohérente, bien que celle-ci soit très importante.

4.4.5 Les traitements effectués sur une base documentaire.

La première tâche est la constitution de la base documentaire, c'est à dire un travail de saisie. La gamme des logiciels Adhoc offrent aujourd'hui de nombreux outils pour faciliter ainsi que contrôler la saisie (e.g. listes d'autorité, qui sont des fichiers en liaison). Une saisie en mode tableur est prévue prochainement, mais ceci dépend plus de la maîtrise de l'interface que de la structure logique de la base.

L'interrogation et la recherche, traitements primordiaux, seront facilités par l'adoption d'une structure proche du modèle conceptuel des données. Il serait possible également d'interroger la base directement en langage SQL ; l'interrogation de ce type est bien sûr réservée au monde des documentalistes qui connaissent la structure de la base et la syntaxe du langage SQL, mais elle permet des recherches complexes, sur plusieurs tables, voire plusieurs bases.

De plus, le "mode point" (écriture de la requête par l'usager) reste pour certains utilisateurs le moyen le plus fiable. Il offre en effet un certain avantage non négligeable vis à vis de l'interface d'interrogation : la recherche doit être pensée et structurée avant que la requête ne soit effectuée.

Il semblerait donc que l'interrogation d'une base possédant une structure directe soit plus rapide, car dans de nombreux cas, la requête portera sur une seule table SQL (alors qu'il y a un minimum d'accès sur trois tables dans la méta-structure actuelle).

Deux traitements sont également à étudier : l'autopostage et l'indexation intégrale, qui existent dans les versions actuelles de la gamme Adhoc.

L'**indexation intégrale** repose sur un principe simple : repérer des "mots", chaînes de caractères délimités par des espaces, virgules, ... Puis constituer un fichier inversé à partir du lexique obtenu.

Néanmoins, l'indexation intégrale comporte plusieurs défauts, entre autres :

- toutes les formes nominales, adjectivales ou verbales d'un terme seront indexées ;

- les syntagmes ainsi que les mots composés seront décomposés, et seuls les unitermes apparaîtront dans la liste.

La mise en oeuvre est simple, mais les résultats de la recherche dans une base importante peuvent être longs à obtenir. Pour faire de la recherche en texte intégral un technique performante d'interrogation, certaines techniques doivent être utilisées :

- couplage avec un thesaurus ;

- extraction des racines des différents termes ;

- etc...

Des techniques d'intelligence artificielle peuvent ainsi rapprocher la recherche en texte intégral d'une interrogation en langage naturel. En effet, l'essor de la gestion électronique de documents (numérisation, reconnaissance, ...) engendrera une part de plus en plus importante du texte intégral dans le monde documentaire, à condition que des outils "intelligents" soient mis à la disposition des gestionnaires d'information.

L'autopostage est le fait d'élargir une recherche par mots-clef à ses génériques, spécifiques ou termes associés (ce qui implique donc l'existence d'un fichier de type thesaurus).

C'est un procédé efficace de recherche dans un fichier lié à un fichier de type thesaurus. Adhoc propose cet outil depuis sa première version (Adhoc Standard), et la hiérarchie est visible graphiquement lors de l'interrogation dans les deux nouvelles versions.

Le principe de l'autopostage est de "descendre" ou de "remonter" la hiérarchie d'un terme puis de constituer une équation de recherche à partir des génériques ou spécifiques (un, plusieurs ou tous) du terme entré.

Le modèle direct pourrait probablement accélérer ce type de traitement en réduisant le nombre d'accès aux tables SQL.

Un point non négligeable du traitement documentaire est l'**impression**, ou la possibilité d'obtenir un document papier, que l'électronique n'a pas encore fait disparaître.

En effet, un écran, d'aussi bonne qualité soit-il, ne remplace pas un document papier. La gamme des logiciels Adhoc a pris ce fait en considération en offrant de nombreuses possibilités d'impressions, et pour des besoins sophistiqués, le module d'impressions (livré avec Adhoc Standard) offrant de multiples possibilités.

Les fiches imprimées peuvent contenir des champs provenant de différents fichiers, des possibilités de récursivité (e.g. impression de thesauri). La mise en page est aisée (interface Windows de type WYSIWYG, "what you see is what you get"), et le logiciel accepte le format RTF (Rich Text Format) développé par Microsoft, qui autorise l'utilisation des différentes polices et styles de caractères (gras, italique, souligné, ...).

L'impression des documents est un aspect à prendre en compte avant la création de la base documentaire. En effet, le document de sortie est une partie du monde réel analysé, (lors de la conception du modèle entité-association, par la méthode MERISE ou autre) et dans certains cas documentaires, c'est un élément primordial.

La conception des entités et des associations doit donc également prendre en compte les résultats désirés en sortie, dont les impressions lorsque celles-ci sont sophistiquées.

5 Conclusion

Ce rapport met mal en évidence les points positifs que m'a apporté ce stage. Il est en effet assez difficile de décrire l'apprentissage d'un logiciel d'OCR ou l'initiation aux réseaux locaux d'un point de vue technique, concernant l'installation et l'utilisation.

Ces points sont pourtant primordiaux : l'enseignement de l'ENSSIB ne peut couvrir l'approche pratique de tous les logiciels et c'est donc bien au sein de l'entreprise d'accueil que cet apprentissage doit être effectué. L'AIO sera pour la première fois présente au SIGED au mois de Septembre 1992, et ma connaissance de ces logiciels me permettra de les seconder lors du salon.

Les tâches qui m'ont été confiées étaient multiples : il n'y a pas eu une construction séquentielle d'un objet, telle qu'une étude de l'existant, avec des méthodes linéaires. Il était initialement prévu une assistance à la conception d'une base documentaire au sein du siège social d'une grande société de construction de Batiments et Travaux Publics, base connectée à un réseau de 2500 stations de travail. Chaque utilisateur devait pouvoir interroger une base "Chantiers" intégrant notices documentaires, images, textes OCéRisés. Or cette mission n'a pas eu lieu pour des raisons internes à l'entreprise utilisatrice.

La partie théorique de mon travail n'a donc pas été validée par cette conception, et c'est pour cette raison que j'ai développé dans ce rapport la théorie des SGBD relationnels, qui est de toute façon un point des plus positifs pour mes connaissances théoriques, et leurs utilisations dans un proche avenir. La quatrième partie de ce rapport, qui est proche de l'objectif initial de ce stage, n'a en effet représentée que 5 à 10% de mon travail de Mars à Juin.

Bibliographie

[Delobel82] Bases de données et systèmes relationnels. / Delobel (C.), Adiba (M.) .- Paris : Bordas, 1982 .- 449 p.

[Desmedt90] DESMEDT (Patrice) .- L'architecture client-serveur sauve les réseaux. *Micro-Systèmes*, Septembre 1990, Num. 111, pp. 114-115.

[Flory87] SQL et DB2. Bases de données. Conception et réalisation. 2° ed. / Flory (A.) .- Paris : Economica, 1987.

[GALACSI86] Les systèmes d'information. Analyse et conception. / GALACSI .- Paris : Bordas, 1986.

[Gardarin89] SGBD Relationnels. Analyse et comparaison des bases de données. / Gardarin (G.), Valduriez (P.) .- Paris : Eyrolles, 1989 .- 403 p.

[Gardarin90] SGBD Avancés. Bases de données objets, déductives, réparties. / Gardarin (G.), Valduriez (P.) .- Paris : Eyrolles, 1990 .- 255 p.

[Koutchouk89] SQL et DB2. Le relationnel et sa pratique. / Koutchouk (M.) .- Paris : Masson, 1989 .- 226 p.

[Pasleau89] Le livre de SQL. / Pasleau (S.) .- Paris : P.S.1, 1989 .- 216 p.

[Poussinov90] POUSSINOV (Nikita) .- Serveur, une transaction s'il vous plait !. *Micro-Systèmes*, Septembre 1990, Num. 111, pp. 117-122.

Annexes

CONVENTIONS SUR LES NOMS DE COLONNE SQL

Les noms de colonnes sont composées en utilisant des abréviations. La liste de ces abréviations est donnée ci-dessous.

VSQLFR Vue Sql sur Fiche Adhoc

BASE	BA
FICHIER	FR
FICHE	FE
CHAMP	CH
SOUS-CHAMP	SC
DONNEE	DO
LIAISON	LI
VUE	VU
INFOVUE	IV
SUITE	SU

LIAISON	LI
MOT-INDEXE	MI
ABREGE	AB
NUMERO	NO
NOMBRE	NB
NOM	NOM
DERNIER	DE
PERE	PE
TYPE	TY
VALEUR	VA
CLE	CLE
INDICATEUR	IR
LONGUEUR	LG
MAXIMUM	MAX
INTITULE	IN
REPETITIF	RF
RECIPROQUE	RE
INDEXE	IX
UNIQUE	UN
CLASSEMENT	CL
ORDRE	OR

exemple : Le dernier numéro de fiche sera codé DE_NO_FE

Pour faciliter l'utilisation est assurer l'unicité des noms de colonnes parmi les tables utilisées, le nom de la colonne sera suivi par le nom de la table . Dans l'exemple ci-dessus, dans le cas d'une donnée associée à un fichier, la forme finale du nom de la colonne sera : DE_NO_FE_FR

STRUTURES DES TABLES ADHOC SOUS SQL

Table BASE :

Cette table contient les informations sur la base elle même. Ces informations sont :

- Le nombre de fichiers de la base,
- Le dernier numéro de fichier,
- Le dernier numéro de champ.

Table FICHIER :

Cette table contient les informations sur les fichiers de la base. Ces informations sont pour chaque fichier :

- Le numéro du fichier,
- Le nom du fichier,
- Le nom abrégé du fichier,
- Le dernier numéro de fiche,
- Le nombre de champs du fichier,
- Un indicateur si le fichier est un dictionnaire.

Table CHAMP :

Cette table contient les informations sur les champs de la base. Ces informations sont pour chaque champ :

- Le numéro du champ,
- Le type du champ,
- Le numéro du fichier auquel appartient le champ (numéro du fichier père),
- Indicateur CHAMP ou SOUS-CHAMP,
- Le numéro du champ père,
- Le nom du champ,
- Le nom abrégé du champ,
- La longueur maximum du champ,
- Le numéro du fichier en liaison,
- Indicateur champ intitulé,
- Indicateur champ équivalent,
- Indicateur champ indéxé,
- Longueur de l'index,
- Indicateur champ unique,
- Indicateur champ repetitif,
- Nombre de répétitions,
- Indicateur champ réciproque,
- Numéro du champ réciproque,
- Indicateur champ mots-indéxés,
- Le numéro du fichier des mots-indéxés s'il il y a lieu (cad : le numéro du fichier dictionnaire pour ce champ)

Table VUE :

Cette table contient les informations sur les vues de la base. Ces informations sont pour chaque vue :

- Le numéro du fichier + le numéro de la vue,
- Le nom de la vue.

Table INFOVUE :

Cette table contient les informations sur les structures des vues de la base. Ces informations sont pour chaque champ :

- Le numéro du fichier + le numéro de la vue,
- Un numéro d'ordre dans la vue,
- Le numéro du champ,
- Le nombre de répétition du champ,
- Mode de classement du champ répétitif,
- Le numéro du champ à afficher.

Table DONNEE :

Cette table contient les données directes de la base. Pour chaque champ de chaque fiche on a :

- Le numéro du fichier ,
- Le numéro de la fiche,
- Le numéro du champ,
- Une clé = numéro du fichier + numéro de la fiche,
- Numéro d'ordre de création pour les champs répétitifs,
- Un indicateur de suite de donnée pour les champs de plus de 200 caractères,
- La valeur du champ.

Table LIAISON :

Cette table contient les données indirectes de la base. Pour chaque champ de chaque fiche on a :

- Une clé = numéro du fichier + numéro de la fiche,
- Le numéro du champ,
- La clé de la fiche en liaison = numéro du fichier + numéro de la fiche.

Description de la table BASE :

Titre : Nombre de fichiers de la base
Nom SQL : NB_FR_BA
Format SQL: CHAR (2)

Titre : Dernier numéro de fichier,
Nom SQL : DE_NO_FR_BA
Format SQL: CHAR (2)

Titre : Dernier numéro de champ
Nom SQL : DE_NO_CH_BA
Format SQL: CHAR (3)

Description de la table FICHIER :

Titre: Numéro du fichier
Nom SQL: NO_FR
Format SQL: CHAR(2)

Titre: Nom du fichier
Nom SQL: NOM_FR
Format SQL: CHAR(15)

Titre: Nom abrégé du fichier
Nom SQL: NOM_AB_FR
Format SQL: CHAR(15)

Titre: Dernier numéro de fiche
Nom SQL: DE_NO_FE_FR
Format SQL: CHAR(7)

Titre: Nombre de champ du fichier
Nom SQL: NB_CH_FR
Format SQL: CHAR(3)

Titre: Type fichier dictionnaire
Nom SQL: TY_MI_FR
Format SQL: CHAR(1)

Description de la table CHAMP :

Titre: Numéro du champ
Nom SQL: NO_CH
Format SQL: CHAR(3)

Titre: Type du champ
Nom SQL: TY_CH
Format SQL: CHAR(1)

Titre: Numéro du fichier père
Nom SQL: NO_FR_PE_CH
Format SQL: CHAR(2)

Titre: Indicateur CHAMP ou SOUS-CHAMP
Nom SQL: IN_CH_SC_CH
Format SQL: CHAR(1)

Titre: Numéro du champ père
Nom SQL: NO_CH_PE_CH
Format SQL: CHAR(3)

Titre: Nom du champ
Nom SQL: NOM_CH
Format SQL: CHAR(15)

Titre: Nom abrégé du champ
Nom SQL: NOM_AB_CH
Format SQL: CHAR(30)

Titre: Longueur maximum champ (limité à 64000)
Nom SQL: LG_MAX_CH
Format SQL: CHAR(5)

Titre: Numéro du fichier en liaison
Nom SQL: NO_FR_LI_CH
Format SQl: CHAR(2)

Titre: Indicateur champ intitulé
Nom SQL: IR_IN_CH
Format SQl: CHAR(1)

Titre: Indicateur champ indéxé
Nom SQL: IR_IX_CH
Format SQl: CHAR(1)

Titre: Longeur de l'index (limité à 200) (compatiblité ADHOC standard)
Nom SQL: LG_IX_CH
Format SQl: CHAR(3)

Titre: Indicateur champ unique
Nom SQL: IR_UN_CH
Format SQl: CHAR(1)

Titre: Indicateur champ répétitif
Nom SQL: IR_RF_CH
Format SQl: CHAR(1)

Titre: Nombre de répétitions (limité à 50)
Nom SQL: NB_RF_CH
Format SQl: CHAR(2)

Titre: Indicateur champ réciproque
Nom SQL: IR_RE_CH
Format SQl: CHAR(1)

Titre: Numéro du champ réciproque
Nom SQL: NO_CH_RE_CH
Format SQl: CHAR(3)

Titre: Indicateur champ mots-indéxés
Nom SQL: IR_MI_CH
Format SQl: CHAR(1)

Titre: Numéro du fichier des mots-indéxés
Nom SQL: NO_FR_MI_CH
Format SQLl: CHAR(2)

Description de la table VUE :

Titre: Clé = Numéro du fichier+Numéro de la vue
Nom SQL: CLE1_VU
Format SQL: CHAR(4)

Titre: Nom de la vue
Nom SQL: NOM_VU
Format SQL: CHAR(15)

Description de la table INFOVUE :

Titre: Clé = Numéro du fichier+Numéro de la vue
Nom SQL: CLE1_IV
Format SQL: CHAR(4)

Titre: Numéro d'ordre dans la vue,
Nom SQL: NO_OR_IV
Format SQL: CHAR(3)

Titre: Numéro du champ
Nom SQL: NO_CH_IV
Format SQL: CHAR(3)

Titre: Nombre de répétition du champ
Nom SQL: NB_RF_CH_IV
Format SQL: CHAR(2)

Titre: Classement du champ (Alphabétique ou Ordre de saisie)
Nom SQL: CL_CH_RF_IV
Format SQL: CHAR(1)

Titre: Numéro du champ à afficher
Nom SQL: NO_CH_AF_IV
Format SQL: CHAR(3)

Description de la table DONNEE :

Titre: Numéro du fichier
Nom SQL: NO_FR_DO
Format SQL: CHAR(2)

Titre: Numéro de la fiche
Nom SQL: NO_FE_DO
Format SQL: CHAR(7)

Titre: Numéro du champ
Nom SQL: NO_CH_DO
Format SQL: CHAR(3)

Titre: Clé = Numéro du fichier+Numéro de la fiche
Nom SQL: CLE1_DO
Format SQL: CHAR(9)

Titre: Numéro d'ordre de création pour les champs répétitifs
Nom SQL: NO_OR_CH_DO
Format SQL: CHAR(2)

Titre: Un indicateur de suite de donnée
Nom SQL: IR_SU_DO
Format SQL: CHAR(1)

Titre: Valeur du champ
Nom SQL: VA_CH_DO
Format SQL: CHAR(200)

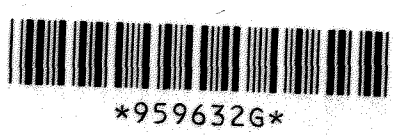
Description de la table LIAISON :

Titre: Clé = Numéro du fichier+Numéro de la fiche
Nom SQL: CLE1_LI
Format SQL: CHAR(9)

Titre: Numéro du champ
Nom SQL: NO_CH_LI
Format SQL: CHAR(3)

Titre: Clé = Numéro du fichier+Numéro de la fiche
Nom SQL: CLE2_LI
Format SQL: CHAR(9)

ENSSIB Bibliothèque Villeurbanne


959632G