ENSSIB
Ecole Nationale Supérieure des Sciences de
l'Information et des Bibliothèques

Université Claude Bernard
Lyon 1

DESS en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage

MIGRATION-FUSION DE BASES DE DONNEES :

REVISION D'UN THESAURUS

Christine Chovet

Sous la direction de
Madame DAVAN

Elf Exploration Production

1997

ENSSIB
Ecole Nationale Supérieure des Sciences de
l'Information et des Bibliothèques

Université Claude Bernard
Lyon 1

DESS en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage

MIGRATION-FUSION DE BASES DE DONNEES :

REVISION D'UN THESAURUS

Christine Chovet

Sous la direction de
Madame DAVAN

Elf Exploration Production

1997

Remerciements

Je tiens à remercier Monsieur CHABIN, Chef du Département Information Documentaire pour l'accueil qu'il m'a réservé au sein de son service, ainsi que ma responsable de stage Madame DAVAN, adjointe au chef du département, pour sa présence aux moments opportuns.

J'adresse toute ma reconnaissance à Madame COSTEMALE, documentaliste principale au SID, pour sa gentillesse et sa disponibilité, ainsi qu'à Monsieur SIGLER, pour son aide et pour m'avoir fait part de son expérience des bases de données internes, tout comme Liliane ECHEGUT, documentaliste,

Je remercie également Mademoiselle VALEYRI, ingénieur documentaliste, pour son soutien et son moral à toute épreuve.

Merci au personnel du SID, pour sa gentillesse, et tout particulièrement à Monsieur JOLIGARD, assistant documentaliste, pour son accueil et son aide précieuse.

Sans oublier les stagiaires : Dominique, Fabienne, Karine, Benoît et tous les autres, sans qui ce stage n'aurait pas été le même.

TITRE :

MIGRATION-FUSION DE BASES DE DONNEES : REVISION D'UN THESAURUS

RESUME :

Cette étude s'inscrit dans le cadre d'une migration-fusion de plusieurs bases de données. Elle s'est plus particulèrement intéressée à la révision d'un thesaurus :

⇨étude de ce thesaurus,

⇨démarches pour le réviser,

⇨analyse des résultats.

Mots-Clés :

révision - thesaurus - mot-clé - migration - fusion - base de données

ABSTRACT :

This study fits into the project of migration-fusion of data bases. The aim was to revise a thesaurus :

⇨study of this thesaurus,

⇨determination of different ways to revise it,

⇨analyse results.

Keywords :

revision - thesaurus - keyword - migration - fusion - data base

SOMMAIRE

PREMIERE PARTIE

PRESENTATION DU GROUPE ELF AQUITAINE

I- LE GROUPE ELF

1- Carte d'Identité du Groupe

En 50 ans, le groupe français Elf est devenu l'un des dix premiers pétroliers mondiaux. De plus, à travers Elf Atochem, Elf est le treizième groupe chimique mondial et, à travers Sanofi, il se classe parmi les trente premiers laboratoires pharmaceutiques mondiaux.

Première entreprise industrielle française avec un chiffre d'affaire de 232,7 milliards de francs et un résultat net de 7 milliards de francs pour l'année 1996, Elf affiche une compétitivité et un professionnalisme de haut niveau. Présent sur les cinq continents, il regroupait, fin 1996, 833 sociétés dans 80 pays. De plus, 85 400 collaborateurs travaillent dans l'une des trois branches du Groupe : hydrocarbures, chimie ou santé. (Fig. 1)

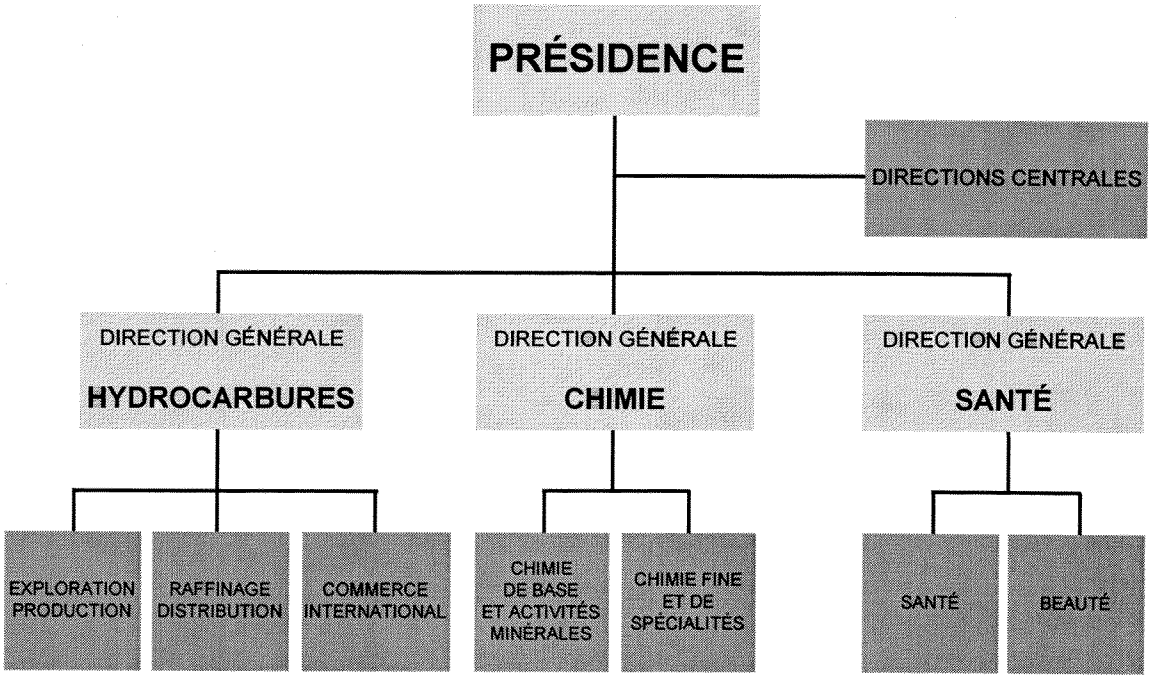


Fig. 1 : Structure du Groupe

L'action Elf est cotée à Paris, New York et sur les principales places boursières européennes. A la Bourse de Paris, Elf Aquitaine est la société française qui a la plus forte capitalisation boursière depuis 1989, et qui, au fil des ans, augmente régulièrement sa distribution de dividendes. De plus, Elf est une des rares entreprises non-américaines admises à la cote du New York Stock Exchange.

2- Historique

L'historique du Groupe est constitué principalement d'une succession de fusions et d'acquisitions.

1939 - Création de la Régie Autonome des Pétroles (RAP), dont l'objet est « la recherche, l'exploration et le transport des hydrocarbures liquides et gazeux en métropole ».

1941 - Création de la Société Nationale des Pétroles d'Aquitaine (SNPA). Associant capitaux publics majoritaires et capitaux privés, sa vocation précise est d'étendre les prospections dans le sud-ouest de la France.

1945 - Mise en place d'une politique de recherche à l'échelle mondiale : création du Bureau de Recherches de Pétrole (BRP), regroupant les petites entreprises de prospection outre-mer.

1951 - Découverte du gisement de Lacq par la SNPA, grâce auquel on produit aujourd'hui soufre, méthane, propane...

1966 - Fusion de RAP et du BRP et naissance de l'ERAP (Entreprise de Recherches et d'Activités Pétrolières), holding publique pour toutes les sociétés existant à l'époque.

1967 - La marque Elf et son emblème, le trépan bleu et rouge, sont dévoilés dans la nuit du 27 au 28 avril.

1970 - Prise de contrôle d'Antar.

1971 - Création d'ATO Chimie, groupement d'intérêt économique paritaire entre Total Chimie et l'Union Chimique Elf Aquitaine (UCEA).

1973 - Premier choc pétrolier. Constitution de SANOFI, holding pour l'animation et le développement du secteur Hygiène-Santé dans lequel le Groupe se diversifie.

1974 - Naissance d'Elf France, issu du groupement des activités de Elf Union et de Elf Distribution.

1976 - Rapprochement de l'ERAP et de la SNPA et création de la SNEA (Société Nationale Elf Aquitaine) qui devient le chef de file du Groupe et qui prend le nom d'Elf Aquitaine.

La SNEA(P) : Société Nationale Elf Aquitaine (Production) filiale à 100% de la SNEA, chargée de la conduite de l'Exploration-Production

1983 - Naissance d'ATOCHEM de la fusion d'ATO Chimie, de Chloé Chimie et de la part de Péchiney Ugine Kuhlmann.

1985 - Création de SANOFI Elf bio-industries par transferts de l'ensemble des activités biotechniques du Groupe à SANOFI.

1986 - Privatisation des entreprises publiques : la participation des actionnaires privés dans le capital de la SNEA passe de 33% à 44% lors de la vente par l'ERAP d'une partie de ses actions Elf Aquitaine.

1991 - Admission du titre Elf Aquitaine à la cote du New York Stock Exchange.

1992 - Nouveau désengagement de l'Etat dans le capital de la SNEA : au terme d'une opération publique de vente portant sur la cession de 5 780 000 actions Elf Aquitaine, la part de l'actionnaire majoritaire est ramenée à 50,8%.

1994 - Privatisation du Groupe en février. La SNEA devient Elf Aquitaine par suite de la cession de l'Etat de sa participation majoritaire dans le capital de la société.

1995 - Développement de la position gazière du Groupe en Europe. Renforcement des positions en Amérique et en Asie. Création d'Elf Aquitaine Gaz.

1996 - Sortie de l'Etat du capital d'Elf Aquitaine avec la cession de ses 10 % résiduels ; et le rachat par Elf de 4,5 % de son propre capital. Création d'Idelfi.

3- Activités du Groupe

Présent de l'exploration à la production, du raffinage à la distribution des hydrocarbures, Elf a su également se développer dans les secteurs de la chimie et de la santé.

3.1- Hydrocarbures

Elf Aquitaine est né des hydrocarbures et réalise, grâce à cette branche, 67% de son chiffre d'affaires, soit 156 milliards de francs en 1996. Les trois grandes divisions de la branche Hydrocarbures sont : Exploration-Production, Raffinage-Distribution, Commerce International et Transports Maritimes.

a- Exploration-Production

Rechercher et exploiter de nouveaux gisements d'hydrocarbures est l'activité première d'Elf Aquitaine grâce à Elf Exploration Production.

Premier prospecteur-opérateur français dans le monde, Elf Aquitaine a acquis cette position par le nombre et la qualité de ses découvertes. Celles-ci sont dues à la maîtrise des techniques les plus élaborées, à l'expérience et au savoir-faire de ses équipes, 11 100 personnes, soit 13% de l'effectif du Groupe en 1996.

b- Raffinage-Distribution

Raffiner et distribuer sont les missions assignées à Elf ANTAR France, fer de lance d'Elf Aquitaine dans ces domaines.

Après la déréglementation du marché pétrolier engagée en 1984-1985 et la libéralisation des prix des produits, Elf Aquitaine a décidé de s'affirmer en

recherchant l'amélioration incessante de sa productivité et en articulant sa stratégie autour d'un maître mot la qualité : qualité des produits, du service-client , du travail et du potentiel humain.

Cet esprit se retrouve dans les quelque 5 300 stations-service, dont 3 000 en France, où est vendue la famille des supercarburants. Ceux-ci sont reconnus par les constructeurs et ont été adoptés par les consommateurs : optane sans plomb 98 et optane sans plomb 95. Il y a également les lubrifiants, 500 produits commercialisés dans plus de 80 pays par Elf Lubrifiants, à travers diverses gammes automobile, industrielle et marine.

c- Commerce International-Transport Maritime

Les modifications intervenues sur les marchés pétroliers au début des années 1980 ont conduit Elf Aquitaine à développer une activité de commerce international de pétrole brut et produits finis (gazole, fiouls, naphta, kérosène,...) ainsi qu'une activité de transports maritimes.

3.2- Chimie

Elf Aquitaine se situe, par sa filiale Elf Atochem, au tout premier rang des chimistes européen, mais également parmi les 15 premiers chimistes mondiaux. Le secteur de la chimie représente 23% du chiffre d'affaire, soit 53.8 milliards de francs en 1996.

Huit centres principaux de recherche, 2 800 chercheurs et de multiples laboratoires composent un puissant outil d'investigation et d'optimisation, dont certaines activités font l'objet de nombreuses cessions de licences sur les cinq continents.

Ce secteur est composé de la chimie de base pour 43% et de la chimie de spécialités pour 57%. La chimie de base, européenne, est articulée autour de trois métiers - pétrochimie, chlorochimie, engrais et chimie minérale - dont les cycles

d'activité se relaient. La chimie de spécialités se compose de la chimie fine et industrielle - thiochimie, chimie du fluor, peroxydes organiques et acryliques - ainsi que de la chimie de spécialités polymères de performance- polymères fluorés, polyamides. Toutes deux sont au niveau mondial et ont des métiers fortement et logiquement complémentaires.

3.3- Santé

Ce domaine représentait 10% du chiffre d'affaires du Groupe en 1996, soit 23,6 milliards de francs.

L'entreprise Sanofi est au second rang français de l'industrie pharmaceutique et parmi les vingt premiers laboratoires mondiaux. Son activité principale est la production de médicaments.

Les parfums et cosmétiques donnent de la couleur et de la magie à la vie : c'est le domaine de la beauté, et avec elle, celui de Sanofi Beauté et des deux affiliés du Groupe, Yves Rocher et Nina Ricci.

Sanofi Beauté recouvre les lignes de parfums et de soins des marques les plus renommées avec notamment Van Cleef & Arpels, Roger & Gallet et Yves Saint Laurent depuis 1993 pour représenter la France et Oscar de la Renta pour les Etats Unis.

3.4- Recherche et Environnement

Elf s'efforce de concilier ses activités industrielles avec le respect de l'environnement. Quelque 7 500 chercheurs représentent le Groupe à travers le monde. Ils oeuvrent dans une vingtaine de centres de recherche, dotés des moyens les plus performants. Ces chercheurs ont pour objectif la mise au point d'innovations ou l'amélioration des gammes déjà existantes. Ils doivent également réaliser des travaux à long terme et des projets à « risque élevé » afin de développer de nouvelles techniques pour mieux protéger notre environnement.

II- ELF EXPLORATION PRODUCTION (EEP)

Depuis mai 1997, Elf Exploration Production est la nouvelle appellation d'Elf Aquitaine Production. EP est donc un secteur de la branche des hydrocarbures.

Elf Exploration Production représente la charnière fonctionnelle entre les différentes filiales Exploration-Production à travers le monde. L'EEP compte quelque 4 700 agents et réalise un chiffre d'affaires à l'exportation de 35 millions de francs par an. La société exerce ses activités dans deux domaines distincts, mais complémentaires :

- **l'ingénierie pétrolière,**
- **la recherche et l'exploitation des hydrocarbures.**

Elf Exploration Production regroupe deux établissements pour le développement de ses activités :

1- Paris où est situé le siège social,

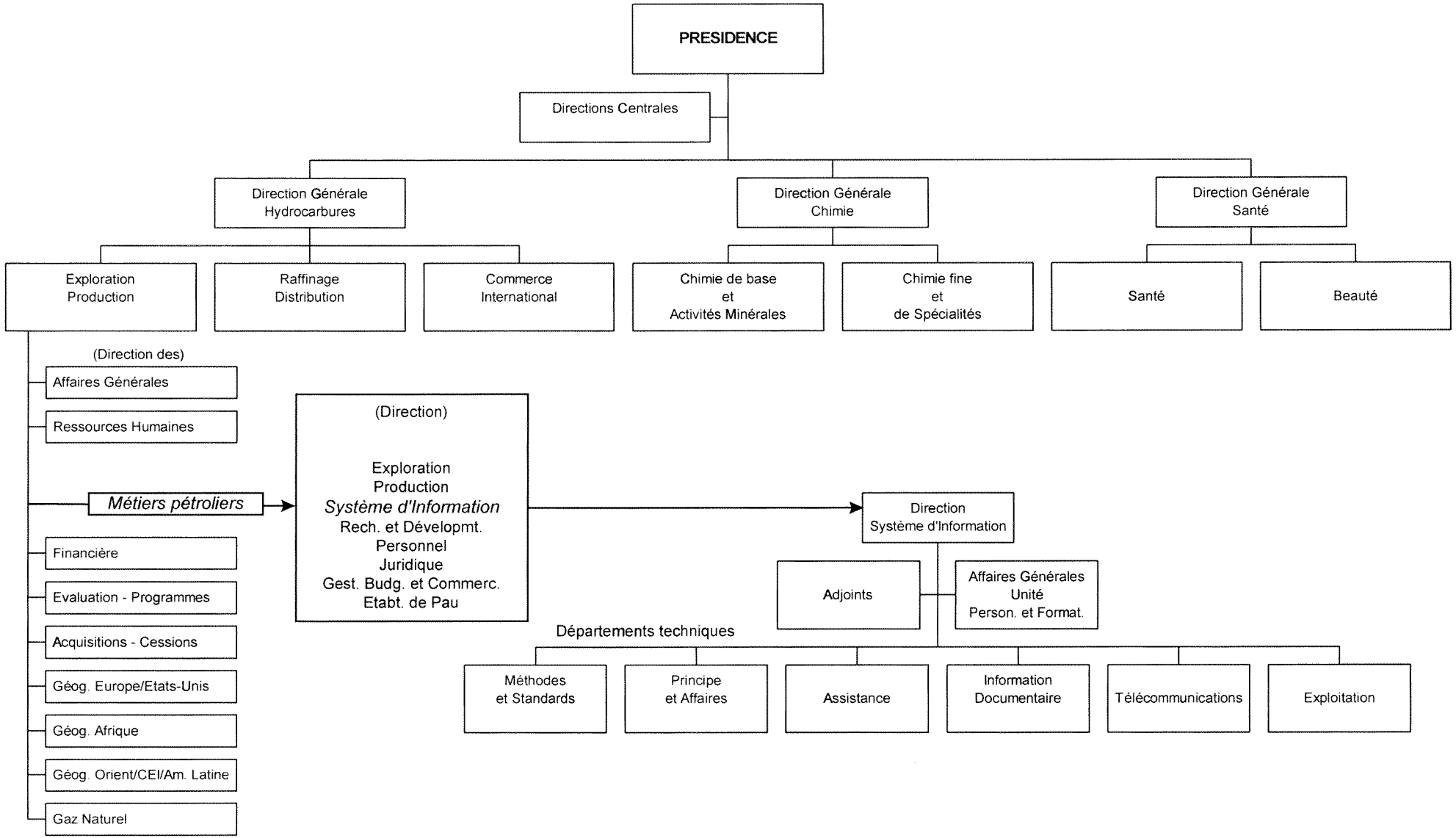
2- Pau où travaillent quelque 2 400 personnes, dont plus de 1 500 ingénieurs et techniciens, entre le *Centre Scientifique et Technique Jean Feger* et le site des *Allées,*

Au 31 décembre 1996, le domaine minier d'Elf Aquitaine couvrait une superficie développée brute de 355 200 km2, se répartissant dans 29 pays. Le Groupe a participé, en qualité d'opérateur, à 33 puits d'exploration et d'appréciation sur un total de 63.

Plus de 60 % de la production de brut revenant à Elf Aquitaine sont originaires du golfe de Guinée ; alors que la quasi totalité des ressources gazières (92 %) provient de la mer du Nord britannique, norvégienne et hollandaise (76 %) et de France (16 %).

Elf Exploration Production a donc une place importante au sein du groupe Elf Aquitaine. Il est constitué de plusieurs directions, dont la *Direction Système d'Information* (DSI), qui me concerne plus directement.

Structure du Groupe ELF - Direction Système d'Information



III- DIRECTION SYSTEME D'INFORMATION (DSI)

J'ai réalisé mon stage au SID (Service Information Documentaire) du CSTJF (Centre Scientifique et Technique Jean Feger). Le SID dépend de l'IDO (département Information Documentaire), qui est l'un des 6 départements de la DSI.

1- Présentation de la DSI

La DSI remplit différentes missions auprès de l'EP grâce à ses 6 départements :

- IDO : information documentaire
- Méthode et standard
- Exploitation des ressources
- Télécommunication
- Assistance informatique
- Projets

☜ voir, ci-contre, **Fig. 2 : Direction Système d'Information**

2- Département Information Documentaire (IDO)

L'IDO (Fig. 3) se charge de gérer l'ensemble des supports de l'information d'Elf Exploration Production et a les fonctions suivantes :

1- conseiller la Direction Exploration Production (DEP) dans la gestion de sa « mémoire » et de son « actif intellectuel » constitués par l'accumulation d'archives et de documents techniques générés par l'activité de l'entreprise,

2- développer, diffuser et rendre plus cohérentes les recherches spécialisées grâce à la présence d'Unités Documentaires Spécialisées (UDS),

3- mettre en œuvre des outils informatiques permettant d'exploiter, de traiter et de gérer un certain volume de données telles que les bases de données exploitées par les UDS,

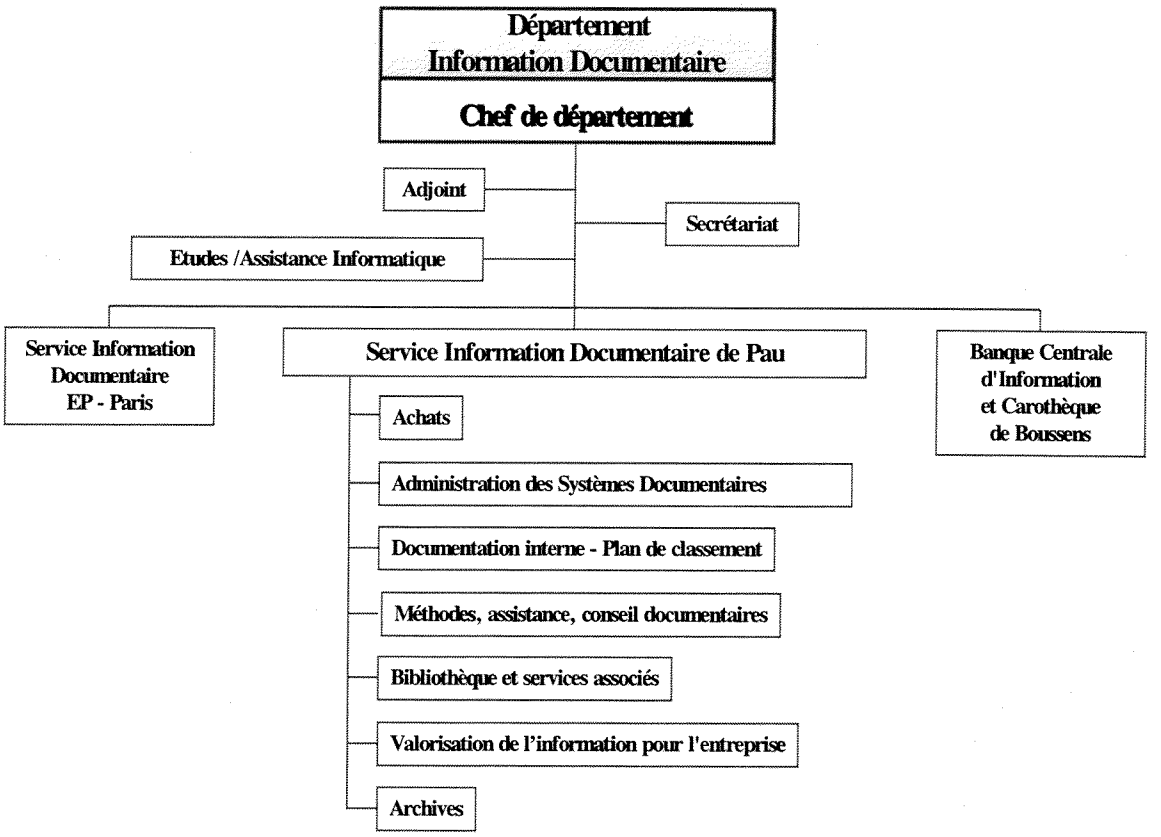
4- rechercher et acquérir, à la demande, des informations aussi bien à l'intérieure qu'à l'extérieure,

5- sensibiliser et former les utilisateurs,

6- développer et consolider le savoir-faire de la DSI ainsi que ses méthodes en information et informatique documentaire,

7- gérer les Services d'Information Documentaire (SID) ainsi que la Banque Centrale d'Information (BCI) de Boussens qui regroupe les archives mortes d'Elf Exploration Production.

Fig. 3 : Organigramme du Département Informatique Documentaire



Trois Services Information Documentaire (SID) composent le Département IDO : Pau, Boussens et Paris.

Le Département IDO à travers le SID gère et distribue la connaissance dans les Métiers de l'Exploration-Production. Il dispose pour ce faire d'une bibliothèque constituée de revues et ouvrages spécialisés, ainsi que des accès à des services extérieurs (serveurs externes, fournisseurs externes) si le document désiré n'est pas disponible sur place. Il met également en place des méthodes de travail destinées à réglementer le métier documentaire : typologie des documents, règles de conservation des documents, règles d'indexation, plan de classement des documents,...

Enfin, le Département IDO fournit des outils de gestion de volumes de données diverses aux UDS, par le biais de données sous-traitées par des sociétés de service informatique, chargées de gérer les documents des spécialités qui les hébergent (exemple de la DEX - Direction EXploration- et de la DPRO - Direction PROduction).

3- Service Information Documentaire (SID)

Le SID reprend les missions citées ci-dessus et il s'occupe principalement des fonctions documentaires suivantes :

- conservation du patrimoine
- recherche de support d'information
- réponse aux questions
- abonnement et acquisition
- bibliographie et circulation de revues
- conseil en organisation et méthodologie documentaire
- formation
- informatique documentaire

- traduction et terminologie

Le SID s'intéresse également aux nouvelles technologies de l'informatique documentaire. Par exemple, à travers son pôle VIE.

VIE signifie Valorisation de l'Information d'Entreprise. Ce pôle intègre :

- la documentation,
- la veille technologique et coneurrentielle, bibliomètrie
- les CD-Roms
- l'internet

Ces nouvelles voies permettent au SID de rester en accord avec l'évolution des technologies et de répondre au mieux aux demandes des utilisateurs, ce qui est, il ne faut pas l'oublier, l'objectif premier de tout service d'information documentaire.

Je vais maintenant vous présenter mon stage de Dess d'Informatique Documentaire, qui s'est déroulé au SID.

DEUXIEME PARTIE

DESCRIPTIF DE L'ETUDE

I- INTRODUCTION

A l'occasion de ce stage chez Elf Exploration Production, j'ai eu l'opportunité de tirer profit de ma double compétence en Géologie et en Informatique Documentaire. En effet, mes études en Sciences de la Terre m'ont permis d'appréhender aisément le vocabulaire utilisé dans les métiers de l'Exploration-Production (EP). Quant à mes connaissances en Informatique Documentaire, leur application fut immédiate. En effet, mon stage s'intégrait dans un ambitieux projet de remaniement des bases de données en place. Pour ma part, je me suis intéressée à la révision d'un thesaurus.

En premier, afin de saisir le « pourquoi » de cette étude, il faut en comprendre le contexte technique. Celui-ci repose sur une idée clé : un accès aux données et aux documents simple et efficace pour l'utilisateur final. Ce dernier désire récupérer vite et bien toutes les données nécessaires à son étude.

C'est pourquoi, Elf Exploration Production est en train de mettre en place un Outil de Manipulation de Données (OMD) pour améliorer l'ergonomie. Ce nouvel outil remplacera les multiples accès aux bases de données existantes ; il sera donc une interface unique entre l'utilisateur final et toutes les bases et en permettra un accès simplifié. Les utilisateurs, en majorité des ingénieurs et des techniciens, pourront accéder plus aisément aux données. Celles-ci sont constituées des données de puits (caractéristiques lithologiques, géométriques,...), des données de réservoir (perméabilité, porosité,...), des données sismiques (profils, log,...), etc.

Les Enjeux de l'OMD :

- **Faciliter et améliorer l'accès aux données patrimoniales**, c'est-à-dire réduction du temps de collecte au démarrage des études, ainsi que du coût et de la durée de celles-ci.
- **Optimiser la gestion des données patrimoniales**, c'est-à-dire réduction des coûts du Système d'Information technique et amélioration de la qualité, de la cohérence et de la pérennité des données.

Parmi les données patrimoniales, il existe différents types de documents. Certains sont accessibles grâce à des références qui sont recensées dans des bases documentaires. Un projet spécifique a été mis en place pour accéder à ces bases documentaires à travers OMD : le projet Doc-Ep.

II- DOC-EP et INFODEP

Le projet Doc-Ep consiste à créer une base documentaire dont le nom est également Doc-Ep. Les utilisateurs pourront accéder aux documents techniques conservés au sein de l'EP grâce à leurs références. Il sera même possible, ultérieurement, d'accéder aux documents primaires. Par exemple, un ingénieur pourra étudier un profil sismique directement sur son écran d'ordinateur, après avoir réalisé, lui-même, sa recherche. Mais n'anticipons pas, Doc-Ep doit d'abord se révéler opérationnel.

1- Contexte

Doc-Ep sera alimentée par Infodep, dans laquelle les mises à jour auront lieu. En effet, Doc-Ep sera une base de données de consultation et Infodep sera la base de données de gestion. Infodep va être créée à partir de trois bases de données documentaires : Infodop, Infodex et Guitar. Une fusion de toutes leurs références est donc nécessaire, mais une difficulté majeure existe : l'hétérogénéité de ces trois bases.

Il s'agit donc d'homogénéiser ces bases de données et de créer une base de données résultante, exploitable par l'OMD. Il y aura ainsi unification du système de gestion des données patrimoniales de subsurface - de sous-sol, en profondeur.

Sur les Fig. 4 et 5, la création de Doc-Ep est symbolisée. Fig.4, les différentes bases de départ sont représentées ainsi que les documents qu'elles recensent. Dans notre étude qui nous concerne, la base Ancrage n'est pas prise en considération (base d'identification des puits de sondage). Puis Fig. 5, Infodep remplace toutes les autres bases. Doc-Ep est consultable par l'OMD et mise à jour par Infodep.

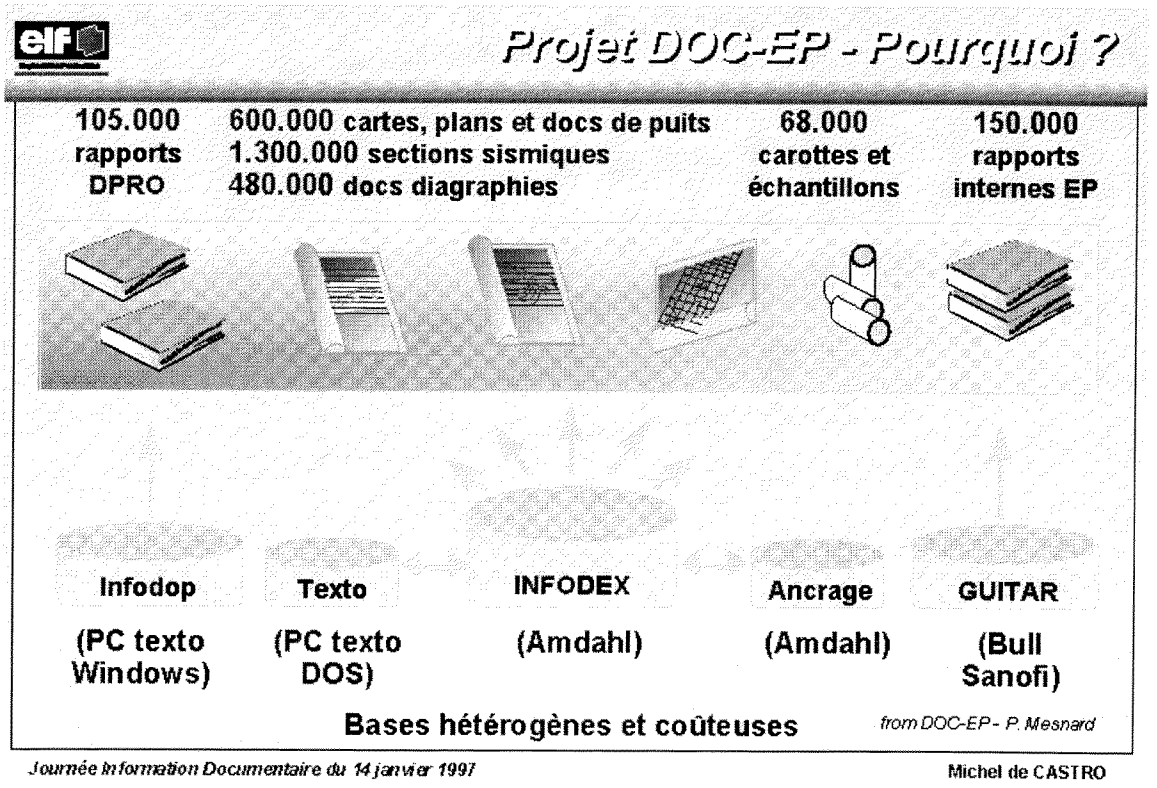


Fig. 4 : les bases documentaires de départ

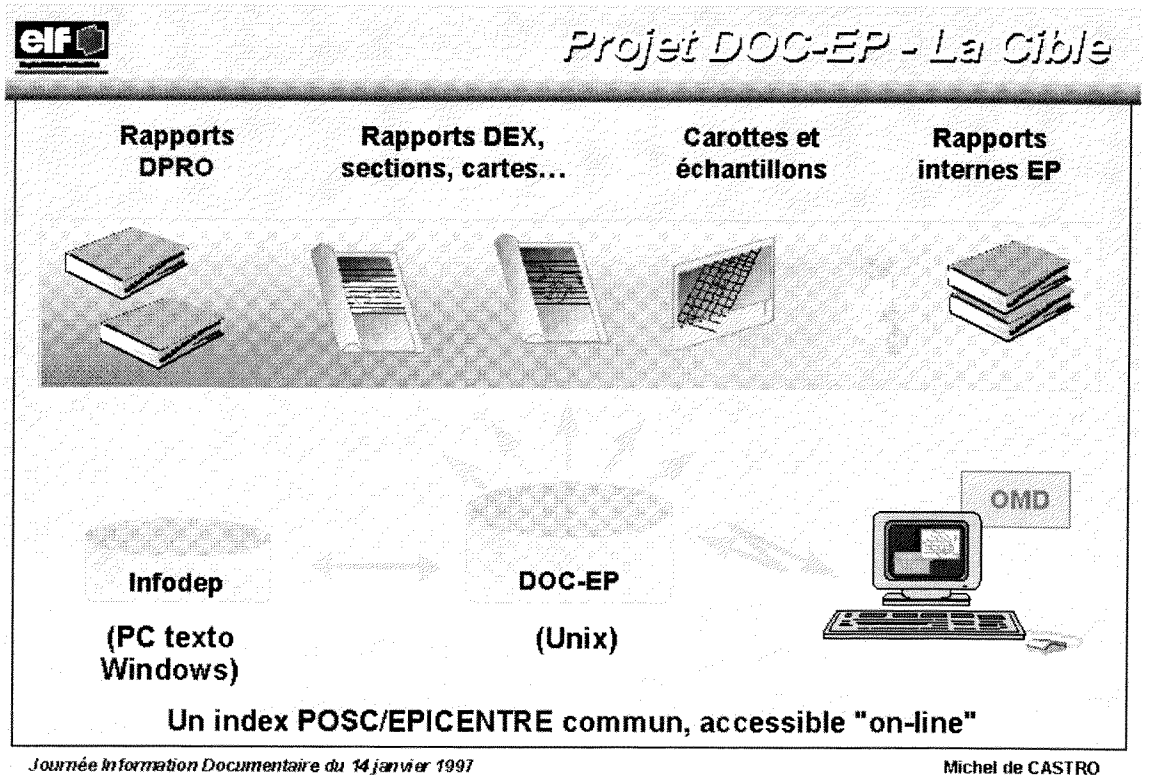


Fig. 5 : Infodep et Doc-Ep

2- Quelques Chiffres

Avant de continuer, il me paraît intéressant d'indiquer quelques chiffres.

Le volume des « références d'objets documentaires » de l'actuel ensemble Infodex+Infodop+Guitar (références de sections sismiques, diagraphies, échantillons, notes et rapports techniques,...) est , doublons exclus, d'environ 2 500 000 références. Ce nombre est en croissance continue.

Sur ce total, le nombre de documents du type « rapport » (domaine spécifique de Doc-Ep) est d'environ 5%, soit environ 125 000. On entend ici par « rapport » un ensemble d'informations diversifiées (par la forme : chiffres, tableaux, graphiques,... ou par la nature : données factuelles, commentées, critiquées, comparées, interprétées,...) dont la caractéristique est d'être « enrobées » dans un texte de forme documentaire classique (ayant un ou plusieurs auteurs, un titre ou un objet, une date, un sommaire, un contenu, une conclusion, un résumé, etc.).

Sur cet ensemble, on peut estimer que le nombre de « vrais rapports de synthèse » (c'est-à-dire qui ne concernent pas exclusivement un seul ou un tout petit nombre d'« objets pétroliers »- ex : sondages) est au moins de 20% (soit environ 25 00 références), c'est-à-dire 1% du volume total des références qui seront gérées par la base patrimoniale technique et accessible par l'OMD. C'est pour exploiter ce « fonds documentaire » de 25 00 références que Doc-Ep a été créée.

Ce chiffre de 1% ne doit pas être considéré comme marginal. Il représente un actif intellectuel considérable. S'en priver temporairement ou totalement serait prendre un risque grave. C'est pourquoi il est nécessaire de réussir la migration-fusion afin que l'utilisateur final ait accès à cette mémoire pour préparer le futur.

3- Description des Trois Bases de Données

a- Infodop

Cette base de données interne gère l'Information Documentaire de la direction Production. Elle recense tous les rapports qui concernent la Production. Ce domaine regroupe les secteurs du forage, du gisement, de la construction, de l'ingénierie,...

Infodop et son thesaurus ont été créés, il y a trois-quatre ans, à partir d'une fédération de petites bases de données. Chacune d'entre elles étaient attachées à un secteur particulier. Par exemple, une de ces bases ne recensait que les rapports liés au secteur du forage. Le thesaurus d'Infodop possède quelques relations hiérarchiques et d'équivalence, mais il s'agit surtout d'une liste alphabétique de termes utilisés dans le domaine de la Production.

b- Infodex

Cette base documentaire recense les documents techniques de l'Exploration, tels que les sections sismiques, les diagraphies, les carottes,...et les rapports. Elle permet d'en gérer les stocks. La direction de l'Exploration englobe tout ce qui concerne la géologie et la géophysique.

Infodex a été créée il y a une quinzaine d'années. Son thesaurus (Fig. 6) est constitué de trois thesaurii principaux :

- le thesaurus type de document - géologie,
- le thesaurus stratigraphique,
- le thesaurus diagraphie opération.

Il ne s'agit pas d'un thesaurus, au sens littéral du terme, mais plutôt d'une liste alphabétique de descripteurs, qui fonctionne par libellé ou par code. Seul le thesaurus stratigraphique est structuré et normalisé par des relations hiérarchiques.

Fig. 6 : Thesaurus Infodex

A	Tables de contrôle : ACQUISITION (ACQ)-MILIEU (MIL)-NATURE DU DOCUMENT (NAT)-DENSITE BANDE (DEN)-FORMAT DONNEES BANDES (FOB)-NOMBRE DE PISTES (NPI)-TYPE DE SONDAGE (STY)-TYPE DE DONNEES BANDE (TYD)
B	Thésaurus : TYPE DE DOCUMENT GEOLOGIE (TD) classé par code
C	Table de contrôle : PRECISION TYPE DE DOCUMENT (PT) documents CP/FA
D	Thésaurus : TYPE DE DOCUMENT GEOPHYSIQUE (TD) classé par code
E	Table de contrôle : PRECISION TYPE DE DOCUMENT (PT) documents SS
F	Thésaurus : INDEXEUR (IND) classé par code
G	Thésaurus : BIBLIOTHEQUE (BIA) classé par code
H	Thésaurus : PAYS (PA) classé par code
I	Thésaurus : PAYS (PA) classé par libellé
J	Thésaurus : ZONE GEOGRAPHIQUE (ZG) classé par code
K	Thésaurus : ZONE GEOGRAPHIQUE (ZG) classé par libellé
L	Thésaurus : PERMIS (PE) classé par code
M	Thésaurus : PERMIS (PE) classé par libellé
N	Thésaurus : SOCIETE (CA-TO-TCO) classé par code
O	Thésaurus : SOCIETE (CA-TO-TCO) classé par libellé
P	Thésaurus : DIAGRAPHIE OPERATION (DOP) classé par code
Q	Thésaurus : DIAGRAPHIE OPERATION (DOP) classé par libellé
R	Thésaurus : STRATIGRAPHIE (STR) classé par code
S	Thésaurus : STRATIGRAPHIE (STR) classé par libellé
T	Thésaurus : ECHELLE (ECH) classé par code
U	Thésaurus : SOURCE (SOU) classé par code
V	Thésaurus : TYPE D'ETUDE (TE) classé par code

c- Guitar

Créée il y a environ trente ans, cette base est la plus ancienne. Elle n'était pas destinée uniquement à l'Exploration-Production (EP), ce qui explique sa dimension et son thesaurus multidisciplinaire. Guitar recense tous les rapports de l'EP ainsi que d'autres secteurs, comme la chimie, pour les documents anciens.

Cette base sera présentée plus explicitement dans le paragraphe III.

4- Difficultés de l'Homogénéisation

La description des bases (cf. II-3) fait émerger un point commun : ces trois bases recensent des rapports de l'EP. Mais l'hétérogénéité de ces bases crée des difficultés pour la migration-fusion de ces rapports.

a- Les documents techniques

Pour un document technique, le problème est tout autre. En effet, des difficultés peuvent être rencontrées lors de la migration si la structure de la nouvelle base diffère de celle des anciennes bases. Il faudra alors établir des tables de correspondance pour effectuer au mieux le transfert. Par contre, le problème de la fusion disparaît puisque les documents techniques sont spécifiques à une base donnée. Il n'y a donc pas de champs communs.

b- Les rapports

Si l'on considère les rapports (pour les références EP), aucune de ces bases n'est exhaustive. En effet, par mesure de sécurité, chaque rapport est indexé dans sa base d'origine (Infodop pour la Production et Infodex pour l'Exploration), puis un double est indexé dans Guitar, qui est la base d'EP. Donc il existe des doublons et même parfois des exemplaires en triple, dus à la fédération d'Infodop. De plus, ces exemplaires n'ont pas été indexés de la même façon suivant la base, car elles s appartenaient, à l'origine, à différentes unités. Malheureusement, ces unités

n'avaient pas toujours les mêmes critères d'indexation. Par exemple, à une époque, les documents ne traitant que d'un puits n'étaient pas indexés sous Guitar.

Aujourd'hui, il nous faut tenir compte de tous ces faits pour réussir la migration-fusion. C'est pourquoi, elle doit être réalisée étape par étape.

c- Démarche choisie

Une gestion en parallèle de ces bases s'avère impossible. Il a donc été décidé, **au niveau de mon étude**, de laisser Infodex de côté. Il ne possède pas de thesaurus à proprement dit, mais un système de tables de codes très perfectionné, comme l'explique le paragraphe II-3-b. Pour le fusionner aux autres, il faudra envisager un système de correspondance.

Les premières étapes seront donc liées à Infodop et Guitar, qui possèdent un thesaurus.

5- Analyse

Dans une base documentaire, il existe de nombreux champs (Fig. 7) qui sont utilisés pour décrire le plus fidèlement un document. Ils peuvent être séparés en deux catégories.

a- Les champs « signalétiques »

Signalétiques, car aucune analyse n'est nécessaire pour remplir ces champs. Il suffit de lire les renseignements sur le document. Ils peuvent donc être considérés avec une difficulté moindre et des processus automatiques peuvent être envisagés afin de normaliser tous ces champs dans la nouvelle base Doc-Ep.

Les champs concernés sont, par exemple, les champs **auteur, titre, date de publication,....** Ils peuvent se présenter différemment d'une base à une autre, mais, par exemple, le nom d'un auteur doit rester le même.

Fig. 7 : Champs d'indexation sous Guitar

OBL	NUM	CODES	LIBELLE DU CHAMP	INDEX
AUTOM	1	ACNR	NUMERO DE REFERENCE OU ACCESSION NUMBER	RANGE
AUTOM	2	DC	DATE DE CREATION	RANGE
AUTOM	3	AM	AUTEUR DE LA MISE A JOUR	INDEX
AUTOM	4	MAJ	DATE MISE MISE A JOUR	RANGE
	5	CS	CODE SECURITE	
	7	NORD	NORD	PHRASE
OBL	19	AU	AUTEUR	INDEX
OBL	20	ORG	ORIGINE DE L AUTEUR	INDEX
	21	DR	DESTINATAIRE DU DOCUMENT	INDEX
OBL	22	TI	TITRE	INDEX
OBL	23	SO	REFERENCES	
OBL	24	DP	DATE DE PUBLICATION	RANGE
	25	PG	PAGES	
OBL	27	LA	LANGUE	THES LF
	28	CI	CODE INDEXEUR	INDEX
OBL	29	NAD	NATURE DU DOCUMENT	THES NF
	30	NAS	NATURE DU SUPPORT	THES KF
	31	IG	IDENTIFIANT GENERIQUE	INDEX
	32	NR	NUMERO DU RAPPORT	
	40	OB	OBSERVATIONS	
	41	NAI	NATURE DE L INFORMATION	INDEX
OBL	42	AV	ARCHIVAGE	INDEX
		LAV	sous champ lieu archivage	THES NF
OBL	45	CC	CODE CLASSIFICATION	THES IF
	47	MP	MOTS CLES PRINCIPAUX	THES
	48	MC	MOTS CLES CONTROLES	THES
	49	ML	MOTS CLES LIBRES	INDEX
	50	AB	RESUME	INDEX
	57	PUI	DESIGNATION NORMALISEE DU PUITS	THES PF
	60	CG	COORDONNEES GEOGRAPHIQUES	INDEX
	70	PRA	CHAMP PRET ALLEES	
		NOMA	NOM	
		DPRA	DATE DE PRET	RANGE
		CPRA	CODE PRET	INDEX
	71	PRE	CHAMP PRET SID CENTRAL	
	74	PRG	CHAMP PRET SID EP PARIS	
	76	PRI	CHAMP PRET ATOCHEM/CRDE	
	77	PRL	CHAMP PRET GRL	
	79	PRP	CHAMP PRET DCGMC	
	80	PRY	CHAMP PRET SOLAIZE	
	81	PRZ	CHAMP PRET BOUSSENS	
	83	PRH	CHAMP PRET ATOCHEM/SIDOC	
	85	PRSA	CHAMP PRET SAD	

b- Les champs « descriptifs »

Descriptifs, car une analyse du document par l'indexeur est obligatoire pour décrire le document à travers ces champs. Nous allons tout particulièrement nous pencher sur le champs des **mots-clés contrôlés** ou **MC**. Ces derniers sont plus difficiles à homogénéiser puisqu'ils ne dépendent pas du même thesaurus.

6- Proposition

La zone des MC est une zone stratégique car elle permet de décrire un concept et un seul. Pour qu'Infodep devienne une base opérationnelle pour les rapports EP, il est nécessaire qu'elle ait un thesaurus unique et exploitable construit à partir des thesaurii existants.

Pour arriver à ce résultat, nous avons choisi le scénario suivant : simplifier chaque thesaurus, puis une fois leur révision terminée, les fusionner. D'autres orientations sont envisageables, mais nous ne les avons pas retenues dans le cadre de mon étude.

Je rappelle que toute la démarche décrite ci-après ne s'applique qu'aux rapports.

III- LE THESAURUS DE GUITAR

Nous allons donc construire un nouveau thesaurus à partir des thesaurii existants et simplifiés d'Infodop et de Guitar. Pour ma part, je me suis penchée plus particulièrement sur la révision du thesaurus de Guitar.

La base Guitar fonctionne sous Basis K sur l'Amdhal. Guitar est une base de données documentaire qui contient les références des rapports émis par différentes entités aussi bien du Groupe que de ses filiales. Les rapports EP représentent 87% du fonds documentaire.

1- Présentation d'Idelfa

En fait, le thesaurus de Guitar, nommé Idelfa, regroupe plusieurs thesaurii. Le thesaurus Idelfa est utilisé pour l'indexation et l'interrogation des banques de données Libra et Guitar. Etant donné la diversité des domaines d'activité des différentes branches du Groupe, ce thesaurus est multidisciplinaire, avec toutefois une orientation plus marquée vers l'exploration et la production pétrolière, ainsi que vers la chimie.

a- Structure d'Idelfa

Idelfa offre 5 types de relation :

- Synonymie : **use**, **uf** (use for), **use...and**, **ufa** (use for...and)
- Abréviation : **ab**, **af** (abrevation for)
- Hiérarchie : - **bt** (broader term=terme générique),
  - **nt** (narrow term=terme spécifique)
- Association : **rt** (related term=terme associé)
- Explication : - **sn** (scope note=note de définition ou d'application),
  - **hn** (historical note=historique)

et contient 5 thesaurii :

- le thesaurus général (Fig. 8) : il est multidisciplinaire et principalement orienté vers les techniques d'Exploration-Production des hydrocarbures et de la chimie. Bien que ce thesaurus contienne de nombreuses relations entre mots, définitions et notes d'application, il est assez peu hiérarchisé ; il s'agit plutôt d'une liste alphabétique.

- le thesaurus géographique (Fig. 9) : il contient des noms de pays, de régions, de bassins sédimentaires, de champs ou de permis pétroliers, ainsi que des divisions administratives locales. Il est hiérarchisé.

- le thesaurus stratigraphique (Fig. 10): il est fortement hiérarchisé. Tous les étages géologiques sont présents ainsi que leurs subdivisions, mais la hiérarchie ne remonte pas jusqu'aux ères Paléozoïque, Mésozoïque ou Cénozoïque.

- le thesaurus des organismes et des sociétés : il n'est pas hiérarchisé. Il comporte de nombreuses notes montrant l'évolution des sociétés dans le temps.

  Ex : ANTAR
  ***hn** Introduit le 1-1-79*
  ***sn** Société constituée le 15 déc.1976*
  Devenue SARL en octobre 1984

- le thesaurus des codes : il comporte trois parties : (1) codes de langues, (2) de classification, et (3) divers.

  Ex : (1) ALLEMAND *(2) GEOLOGIE* *(3) NORVEGE*
  ***ab** ALL* ***use** 08A* ***use** WW*
  D
  DE
  GER

```
PETROGENESE
  UF  GENESE ROCHE
  SN  PRECISER EVENTUELLEMENT LA NATURE
      DE LA ROCHE. POUR LE PETROLE,
      UTILISER GENESE HYDROCARBURE
  RT  DIAGENESE

PETROGRAPHIE
  UF  LITHOLOGIE

PETROLE BRUT
  UF  HUILE
  RT  BRUT AROMATIQUE
      BRUT COMMERCIAL
      BRUT LEGER
      BRUT LOURD
      BRUT NAPHTENIQUE
      BRUT PARAFFINIQUE
      BRUT REFERENCE
      BRUT SYNTHETIQUE
      BRUT TRAITE
      MELANGE BRUT
```

Fig. 8 : extrait du thesaurus général

```
FALKLAND
  BT  AMERIQUE SUD
  UF  MALOUINES

FERNANDO DE NORONHA
  BT  AMERIQUE SUD
      BRESIL

FERNANDO PO
  BT  AFRIQUE
      GUINEE EQUATORIALE

FEROE
  BT  DANEMARK
      EUROPE

FIDJI
  BT  OCEANIE

FINISTERE
  BT  FRANCE

FINLANDE
  BT  EUROPE
```

Fig. 9 : extrait du thesaurus géographique

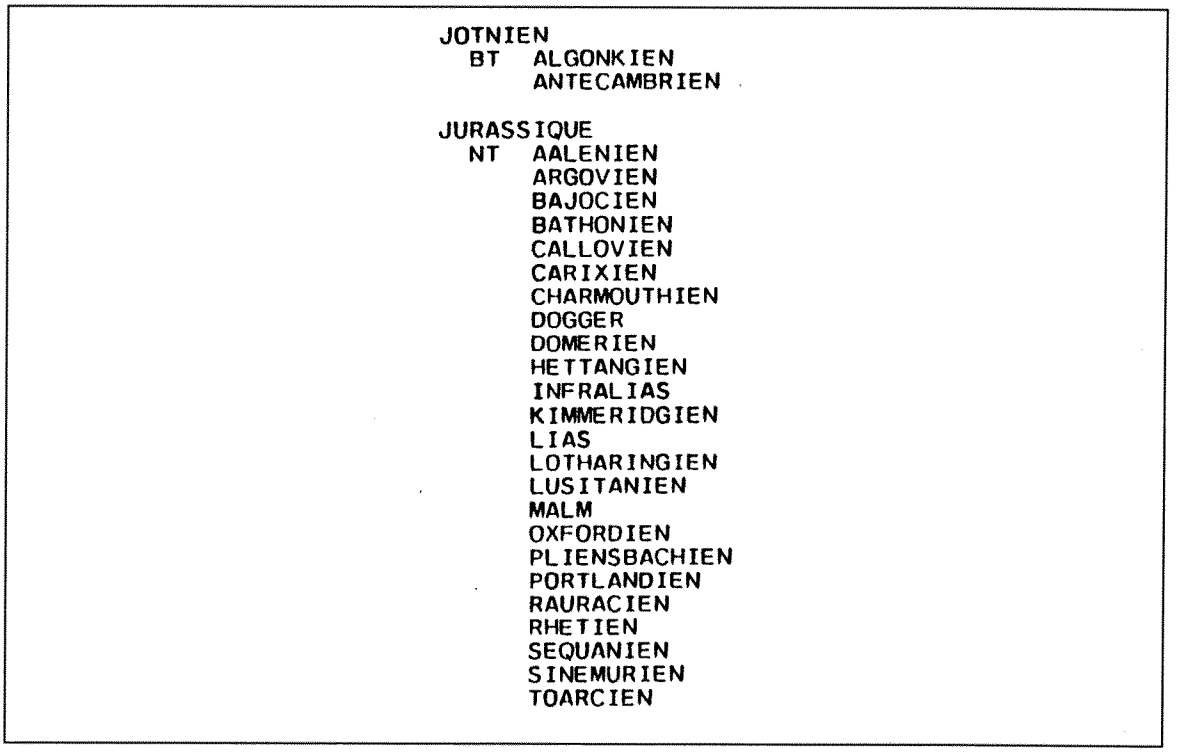


Fig. 10 : extrait du thesaurus stratigraphique d'Idelfa

b- Utilisation d'Idelfa

Lorsqu'il a été créé, il y a environ trente ans, le thesaurus Idelfa était destiné à l'ensemble du Groupe. Puis, il a été spécifiquement utilisé pour indexer sous Guitar et sous Libra.

Libra est une base documentaire, qui permet de gérer les livres de la bibliothèque. Son thesaurus a donc besoin d'être multidisciplinaire et Idelfa correspond très bien à cette demande. C'est pourquoi Libra n'est pas intégrée dans le projet de migration-fusion de Doc-Ep. Par contre, Guitar est une base de rapports internes spécifiques à l'EP et Idelfa ne lui correspond plus.

2- Révision d'un Thesaurus

a- Qu'est-ce qu'un thesaurus ?

« Un thesaurus est, avant tout, un lexique de mots autorisés pour une zone donnée » (L.Sigler). Ces mots sont appelés mots-clés contrôlés ou MC. *« L'utilité majeure d'un thesaurus est de normaliser le vocabulaire utilisé lors de l'indexation et de l'interrogation, afin d'améliorer les performances de la base de données »* (L.Sigler).

b- Comment constituer un thesaurus de MC ?

« La création d'un thesaurus de mots-clés passe par l'étape suivante :
Etablir une liste de tous les mots-clés dont on estime avoir besoin (aussi définitive que possible, les modifications ultérieures, ajouts mis à part, devant être réduites au minimum), en respectant ces deux règles essentielles :

1- Un concept doit être décrit par un mot-clé et un seul (élimination de synonymes ou bien prise en compte automatique de ceux-ci).
2- Un mot-clé ne doit avoir qu'un seul sens. » (L.Sigler).

Notre cas est différent puisque nous avons déjà un thesaurus. Mais, il paraît important de rappeler les principes fondamentaux de création d'un thesaurus. En effet, les avoir en mémoire nous permettra de réviser notre thesaurus au mieux afin de « créer » le futur thesaurus d'Infodep.

c- Les mots-clés ou MC

L'étude des MC a donc un rôle important dans la révision du thesaurus de Guitar.

La fréquence d'utilisation des MC reflète leur degré de pertinence. Si personne ne les utilise, on peut envisager la possibilité qu'ils soient superflus.

Les rapports ne constituent que 1% du volume total des références qui seront traitées par l'OMD (cf. II-2). Le thesaurus d'Infodep doit être dédié à l'Exploration-Production ; il convient donc d'éliminer les MC inutiles.

d- Première conclusion

Pour réviser un thesaurus, la meilleure méthode semble donc d'éliminer des MC. Maintenant une question se pose :

quels critères retenir pour éliminer un mot-clé d'un thesaurus ?

Nous allons essayer d'y répondre en expérimentant deux méthodes. La première est dite « pragmatique » et la deuxième est plus « réfléchie ».

3- Méthode « Pragmatique »

Cette méthode repose sur une analyse de la fréquence d'utilisation des mots-clés dans l'index.

Une fréquence élevée signifiera que le terme est pertinent, puisque l'indexeur l'utilise souvent. A l'inverse, une fréquence basse correspondra à un terme peu utilisé, donc que l'on peut juger « superflu ».

De ce point de vue, la révision d'un thesaurus est pragmatique. Il suffit de fixer le nombre de MC que l'on désire garder et on élimine tous les autres suivant une fréquence d'utilisation décroissante lors de l'indexation.

Cette méthode extrême est simple et efficace, peut-être même trop. En effet, est-il vraiment pertinent d'éliminer tous ces MC sous prétexte que les indexeurs les utilisent peu ? Avant tout jugement hâtif, il convient de présenter la deuxième méthode.

4- Méthode « Réfléchie »

Cette méthode se fonde également sur une analyse de la fréquence d'utilisation des MC, mais cette fois lors de l'interrogation. Basis K, sous lequel fonctionne Guitar, permet d'avoir accès au « monitor » de Guitar. Celui-ci est un fichier où sont enregistrées les interrogations.

Ainsi, les mots interrogés le sont à la demande du personnel EP qui est le futur utilisateur de DOC-EP. Cela signifie donc, de ce point de vue, que les termes les plus fréquents sont les plus pertinents. Il faut donc les conserver.

5- Réflexions et Objectifs

En fait, il semble difficile de déterminer laquelle de ces deux méthodes est la meilleure. Qu'est-ce qui est le plus important ? Les mots-clés utilisés pour l'indexation, qui sont un reflet du patrimoine de la société au cours du temps ou ceux utilisés lors de l'interrogation, qui indiquent les besoins réels de l'entreprise ?

Une révision est nécessaire. Le mieux est donc de comparer ces deux listes de fréquences afin d'observer s'il existe une corrélation.

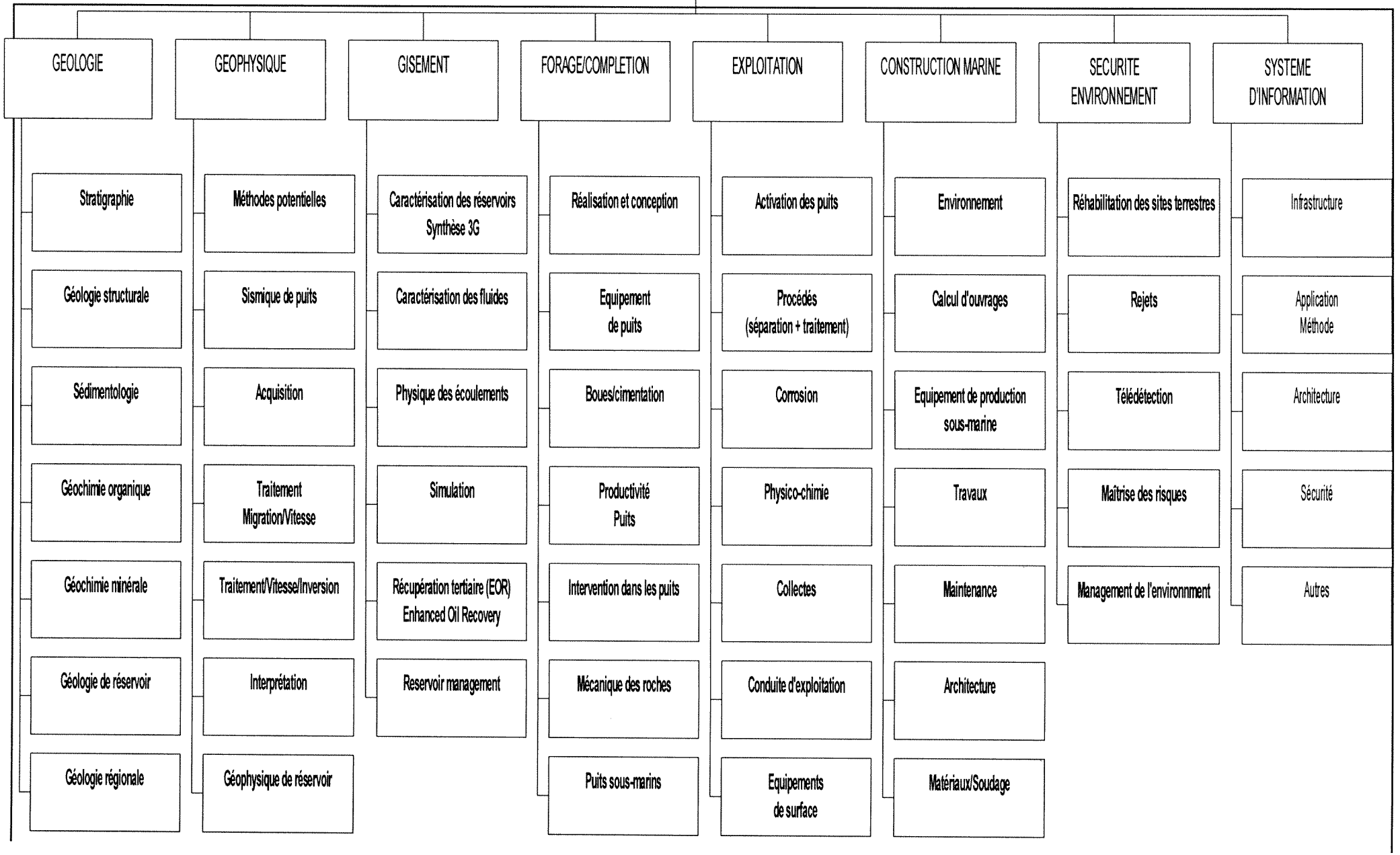
En effet, si la corrélation est forte entre les fréquences d'apparition des mots-clés, cela signifiera que les deux méthodes sont équivalentes. Dans ce cas, on utilisera la première, plus simple à appliquer. Par contre, si la concordance s'avère mauvaise, une solution moins radicale devra être trouvée.

Le cadre de mon étude ayant été défini ainsi que l'objectif poursuivi, nous allons maintenant nous intéresser à la partie pratique de mon stage :

comparaison des listes de fréquences et analyse des résultats obtenus.

elf Exploration Production

ARBRE DES DOMAINES



IV- ETUDE

Après cette approche conceptuelle de mon étude, abordons maintenant l'aspect pratique de mon stage et les problèmes rencontrés. Je tiens à signaler qu'il est essentiel de garder à l'esprit la démarche théorique, car les deux parties sont interdépendantes. En effet, j'ai mené de front les deux aspects, tout au long de mon stage, dans le but d'obtenir des résultats probants.

1- Introduction

A mon arrivée, l'étude proposée était certes passionnante, mais trop ample. Il s'agissait alors de rédiger le manuel d'une démarche qualité concernant la migration-fusion de données documentaires extraites de bases de données anciennes. Plus spécifiquement, je devais traiter 25 000 références provenant de trois bases différentes (cf. II-2).

Ce projet demandait à être précisé. Aussi, fut-il décidé que mon étude porterait sur la base Guitar (cf. II-3-c et III) afin de me fournir un point de départ concret pour mon étude.

2- Premiers Pas

Tout d'abord, je me suis familiarisée avec la base Guitar. Après avoir consulté les manuels d'utilisation, j'ai procédé à de nombreuses interrogations, exploitant ainsi l'Arbre des Domaines (Fig. 11 : voir page ci-contre). Cet arbre a été validé par les spécialistes de l'Exploration-Production. Il est une représentation de la connaissance de l'EP. Il semble alors pertinent de l'utiliser comme référentiel puisqu'il est reconnu par les professionnels, futurs utilisateurs de Doc-Ep.

J'ai interrogé Guitar en utilisant les domaines et sous-domaines comme **mots principaux** (MP), **mots-elés contrôlés** (MC), **mots libres** (ML), puis comme appartenant au **titre** (TI) ou au **résumé** (AB) et à l'**index de base** (BI).

Les interrogations successives ont permis de constituer des tableaux résultats (Tab. 1). Il semblait intéressant de croiser les différents termes afin de tester une possible redondance des descripteurs lors de l'indexation. Mais la multiplicité des termes et le nombre des documents ont fait que cette démarche ne saurait, dans un premier temps, aboutir concrètement à un résultat.

La complexité de l'analyse révèle la difficulté d'appliquer une démarche qualité.

3- Une Approche Différente

Ainsi, mon étude fut réorientée et évolua vers une démarche plus concrète : la révision du thesaurus de Guitar en vue d'une migration-fusion.

En effet, au lieu de m'appuyer sur l'Arbre des Domaines, certes reconnu par des professionnels, mais encore non-utilisé comme référentiel pour l'indexation ou l'interrogation de Guitar, il fut jugé plus adéquat que j'exploite directement le fichier des mots-clés contrôlés de Guitar.

P.S. Cette étape est liée à la démarche théorique développée précédemment (cf. III).

4- Fichier des Mots-Clés Contrôlés liés à l'Indexation (cf. Schéma 1)

Ce fichier, à l'état brut, se présentait sous la forme de deux colonnes en format texte. L'une contenait les MC et l'autre, la fréquence qui leur était associée dans l'index. J'avais donc là, un premier élément pour juger de l'utilisation des MC. Mais avant que ce fichier ne soit exploitable, quelques transformations étaient nécessaires.

En effet, après l'ouverture du fichier sous Excel, un **problème** est apparu : la **relation de synonymie USE**. Sa présence rendait le tableau difficilement manipulable pour des comparaisons.

Ex : houille USE charbon

Houille est le terme dit rejeté et charbon le terme dit préféré.

		BI	MP	MC	ML	TI	AB
I-	GEOLOGIE*	12913	5922	12077	78	2293	892
1-	stratigraphie*	29871	1276	29686	117	1038	628
	*stratigraphie**	*5967*	*150*	*5554*	*0*	*27*	*98*
2-	geologie*structurale*	28	0	0	28	0	0
	*geologie*structurale**	*28*			*28*		
	structurale*	7472	0	0	3	1090	1017
	*structurale**	*1599*			*0*	*87*	*105*
3-	sedimentologie*	6420	1305	6308	0	510	155
	*sedimentologie **	*993*	*35*	*834*		*22*	*23*
4-	geochimie*	7381	1248	7239	15	1125	580
	*geochimie**	812	39	709	0	53	63
5-	geochimie*organique*	4185	654	4185	0	1	0
	*geochimie*organique**	*213*	*13*	*173*		*0*	
6-	geochimie*minerale*	1746	184	1746	0	1	0
	*geochimie*minerale**	*274*	*2*	*259*		*0*	
7-	geologie*reservoir*	1	0	0	1	0	0
	*geologie*reservoir**	*1*			*1*		
	reservoir*	27196	1204	21990	106	5099	4345
	*reservoir**	*2800*	*24*	*2170*	*0*	*98*	*125*
8-	geologie*regionale*	13	0	0	13	0	0
	*geologie*regionale**	*13*			*13*		
	regionale*	285	0	0	0	269	272
	*regionale**	*86*				*41*	*77*

est nul implicitement car l'un des deux termes croisés a donné un résultat nul

Légende du tableau :

Comme exemple, la branche GEOLOGIE avec ses sous-domaines (2 à 8) a été choisie dans l'Arbre des Domaines (cf. Fig. 11).

Chaque terme a été interrogé en tant que BI, MP, MC, ML, TI et AB. Les résultats de ces interrogations apparaissent dans le tableau. Les lignes en italique correspondent à une interrogation du domaine principal (I), ici GEOLOGIE, avec un sous-domaine (2 à 8).

Ex : find mc=geologie and mc=stratigraphie.

L'astérisque * correspond à un « stem » : le terme est cherché dans des associations.

*Ex : **geologie*** correspond à **geologie**, **geologie appliquee**, **geologie miniere**, **geologie petroliere** et **geologie terrain**. (les accents ne sont pas reconnus dans Guitar)*

Tab. 1 : Résultats d'interrogation de Guitar

J'ai donc réalisé la démarche suivante :

a- j'ai vérifié que chaque mot « préféré » soit présent dans le fichier

- s'il l'était, les lignes contenant le mot rejeté et le « use » étaient effacées,
- s'il ne l'était pas, seule la ligne du mot rejeté disparaissait et le mot préféré prenait sa fréquence en s'insérant dans le tableau.

b- à la fin, un tri alphabétique a été réalisé afin de réorganiser le tableau.

Synthétisée en quelques lignes, cette démarche semble facile à appliquer. En fait, la difficulté principale est liée à la taille du tableau (Tab. 2).

	Tableau « brut »	Tableau « final »
Nombre de lignes	8062	5567

Tab. 2

Il y a donc eu élimination de 2495 lignes. Cette manipulation a été réalisée grâce à une macro que j'ai créée. Celle-ci est décrite ci-après.

MACRO :

a- création d'une colonne de numérotation des lignes, qui permet de connaître la position d'un terme. Copie du tableau.

b- classement alphabétique pour mettre en évidence les termes préférés qui sont redondants.

c- comparaison de chaque mot avec la « fréquence » USE au terme précédent et au suivant :

- si le mot est en double, la ligne contenant le USE est effacée ainsi que la ligne contenant le mot rejeté qui lui est associé. Ceci est possible grâce à la numérotation des lignes. Le terme rejeté se trouve à la ligne n-1.

- si le mot n'est pas redondant, seule la ligne avec le terme rejeté est éliminée et sa fréquence remplace USE.

d- cette opération terminée, un tri alphabétique est réalisé ainsi qu'une nouvelle numérotation des lignes. Le fichier est maintenant manipulable.

5- Fichier de Mots-Clés Contrôlés liés à l'Interrogation (cf. Schéma 2)

Ce fichier a été plus difficile à obtenir puisqu'il a fallu le créer. On m'a fourni des fichiers bruts du monitor de Basis K concernant les interrogations sous Guitar.

Pour que l'étude soit pertinente, une évaluation sur une durée de 6 mois (janvier à juin 1997) a été décidée. En effet, un mois ne fournissait pas assez de mots pour que l'étude ait une quelconque valeur.

Comme pour le fichier des MC d'indexation, ces fichiers doivent être traités afin d'être exploitables. Tout d'abord, ils sont autant le reflet de toutes les interrogations pertinentes que des erreurs de frappe. De plus, les interrogations portent sur de nombreux champs, or seuls certains d'entre eux m'intéressaient. En effet, je voulais obtenir les mots-clés contrôlés interrogés.

Voici qu'elle a été ma démarche :

a- récupération de tous les BI et MC. J'entends par là, toutes les questions comportant « BI= » et « MC= ». En effet, lorsque l'on interroge BI (Basic Index), on cherche implicitement dans le Titre, les Mots Principaux, les Mots-Clés Contrôlés et les Mots Libres. Il faut donc tenir compte de ce champs.

Bien sûr, cela crée du bruit, mais ignorer BI, c'est négliger de nombreux MC. Les documentalistes, qui utilisent régulièrement Guitar, m'ont expliqué qu'elles interrogeaient souvent les MC à travers BI. De plus, BI est le champs qui est appliqué par défaut.

Ex : find GEOLOGIE correspond à find BI=GEOLOGIE

b- obtention de 6 fichiers Excel contenant une colonne avec BI et l'autre avec MC.

c- élimination des préfixes « BI= » et de « MC= »

d- union des deux colonnes

e- tri alphabétique

f- une macro m'a aidée pour réaliser l'étape suivante. Elle dénombrait la fréquence d'apparition d'un terme, puis la copiait dans une nouvelle colonne en face du terme correspondant.

g- fusion des 6 fichiers ainsi traités et tri alphabétique, puis de nouveau utilisation d'une macro. Comme la précédente, elle cherche les mots semblables, mais cette fois, elle additionne leur fréquence dans une nouvelle colonne en face du terme correspondant.

J'ai donc un fichier de mots-clés pertinents et manipulable.

6- Etape Finale avant la Comparaison

Nous avons donc deux fichiers exploitables, mais ils n'ont pas encore leur forme définitive. En effet, pour obtenir les MC qui nous intéressent réellement, un nouveau traitement est nécessaire.

Celui-ci comprend l'élimination des MC de thesaurus géographique et ceux du thesaurus stratigraphique. En effet, ceux-ci ne permettent pas de décrire un savoir-faire, une synthèse ou une recherche. Ils donnent une information précise, mais qui ne peut évoluer. Le thesaurus stratigraphique constitue une référence en géologie et ne peut donc pas être transformé. Le thesaurus géographique, quant à lui, s'adaptera aux évolutions politiques (ex : l'URSS a disparu). On aura toujours besoin de ces MC pour définir un lieu géographique ou une couche stratigraphique, mais ni l'un ni l'autre ne sont réellement caractéristiques de l'EP.

Ex : rapport concernant une couche géologique de l'Albien au Vénézuela.

Les **MC du thesaurus général** sont ceux qui **nous intéressent**. Ils permettent de définir un savoir-faire, une synthèse ou une recherche, ce qui représentent 20% des données de l'EP. Pour les obtenir, j'ai utilisé une base Access dans laquelle j'ai développé la macro suivante.

MACRO :

1- importation du thesaurus géographique dans la table ThGeo

2- importation du thesaurus stratigraphique dans la table ThStr

3- importation du thesaurus général dans la table ThG

4- importation du fichier à traiter dans la table T0

5- comparaison de T0 et ThGeo grâce à la requête 1

6- création de deux fichiers :

- T1 contient les MC différents de ThGeo
- T1b contient les MC présents dans ThGeo

7- comparaison de T1 et ThStr grâce à la requête 2

8- création de deux fichiers :

- T2 contient les MC différents de ThStr
- T2b contient les MC présents dans ThStr

9- comparaison de T2 et ThG grâce à la requête 3

10- création de deux fichiers :

- T3 contient les MC **présents** dans ThG
- T3b contient les MC différents de ThG

T3 contient les MC qui vont être comparés.

Cette macro sera exécutée deux fois (Tab. 3 et 4)) afin d'obtenir les deux fichiers qui nous intéressent :

- avec le fichier des MC de l'index (**MC index**) afin d'éliminer les MC géographiques et stratigraphiques, ainsi que les MC qui n'appartiennent pas au thesaurus général,

MC index	Nombre de MC dans les tables					
à l'origine	T0	*5567*				
après ThGeo	T1	*3083*	T1b	*2484*	T1+T1b=T0	*5567*
après ThStr	T2	*2635*	T2b	*448*	T2+T2b=T1	*3083*
après ThG	**T3**	***2294***	T3b	*341*	T3+T3b=T2	*2635*

Tab. 3 : Visualisation étape par étape de l'exécution de la macro pour MC index

- avec le fichier des MC du monitor (**MC monitor**) pour éliminer les MC géographiques et stratigraphiques, et le bruit créé par la sélection de BI (cf. IV-5).

MC monitor	Nombre de MC dans les tables					
à l'origine	T0	*1264*				
après ThGeo	T1	*1095*	T1b	*169*	T1+T1b=T0	*1264*
après ThStr	T2	*1083*	T2b	*12*	T2+T2b=T1	*1095*
après ThG	**T3**	***126***	T3b	*957*	T3+T3b=T2	*1083*

Tab. 4 : Visualisation étape par étape de l'exécution de la macro pour MC monitor

Avant d'analyser le résultat de la comparaison, examinons deux points de réflexion (Tab. 5) :

- Après l'exécution de la macro, il subsiste environ 50% de MC index et seulement 10% de MC monitor. Ces 10% démontrent un bruit important, dû à la sélection de BI. En effet, le tableau 4 prouve qu'il ne s'agit pas de MC géographiques et stratigraphiques. Il faudrait donc refaire cette étude en sélectionnant uniquement les MC dans le fichier origine.

- Par contre , il est intéressant de souligner que, *sur une durée de 6 mois*, seuls *126 MC du thesaurus général* ont été utilisés lors d'interrogations. Ce résultat vient donc conforter l'idée que ***l'EP a besoin d'un thesaurus réduit et plus spécifique.***

	Nombre de lignes dans T0	Nombre de lignes dans T3	T3 = x% de T0
MC index	5567	**2294**	~50%
MC monitor	1264	**126**	~10%

Tab. 5 : Récapitulatif

7- Comparaison et Analyse

Nos deux fichiers sont maintenant composés des MC que l'on désire comparer. Comme hypothèse de travail, MC index servira de référentiel puisqu'il a le plus grand nombre de MC (Tab. 5).

Voici un extrait (Tab. 6) du tableau résultat de la comparaison.

Index		Mots-Clés	Monitor	
occurence	rang		rang	occurence
48096	**1**	*SONDAGE PETROLIER*		
29404	**2**	STRATIGRAPHIE	**104**	*3*
18588	**3**	PETROGRAPHIE	**119**	*1*
18579	**4**	**INTERPRETATION**	**5**	*49*
18110	**5**	*RESERVOIR GEOLOGIQUE*		
13046	**6**	SISMIQUE REFLEXION	**85**	*4*
12015	**7**	*CAROTTE*		
11756	**8**	*EXPLORATION PETROLIERE*		
10695	**9**	**PETROLE BRUT**	**35**	*12*
10358	**10**	MISE EN ŒUVRE GEOPHYSIQUE	**76**	*4*
10109	**11**	**PRESSION**	**7**	*34*
9180	**12**	*ESSAI PRODUCTION*		
8466	**13**	**FORAGE**	**11**	*24*
8311	**14**	*GAZ NATUREL*		
8269	**15**	**SISMIQUE**	**1**	*102*
8179	**16**	**DIAGRAPHIE**	**4**	*53*
7563	**17**	*CARACTERISTIQUE RESERVOIR*		
7436	**18**	SISMIQUE PUITS	**123**	*1*
7319	**19**	PROJET	**49**	*8*
7214	**20**	*VITESSE SISMIQUE*		
7039	**21**	**TECTONIQUE**	**16**	*19*
6791	**22**	BOUE FORAGE	**109**	*2*
6705	**23**	**CORRELATION**	**30**	*12*
6506	**24**	**CARTE**	**39**	*9*
6315	**25**	**SEDIMENTOLOGIE**	**24**	*15*
6268	***26***	*EXPLOITATION PETROLIERE*		
6185	***27***	*MICROPALEONTOLOGIE*		
6120	***28***	*MARQUEUR GEOPHYSIQUE*		
6119	***29***	*GEOLOGIE PETROLIERE*		
5998	***30***	*EVALUATION PROSPECT*		

Tab. 6 : Les 30 premières lignes du tableau comparatif

Légende du tableau :

a- dans les colonnes rangs : le rang est lié à la fréquence

Ex : le rang 1 correspond à la fréqence la plus élevée

b- dans la colonne Mots-Clés :

- **en gras : MC « corrélés », qui ont un rang élevé dans les deux listes**
- *en italique : MC non-présents dans le monitor*
- normal : MC qui ont un rang opposé dans les deux listes

a- Analyse du tableau n°6

Ces trente premières lignes nous permettent déjà de mettre en relief quelques points :

- 10 MC (écrits en gras) ont vraiment une bonne corrélation. En effet, ils ont une fréquence d'apparition élevée dans les deux listes,
- 7 MC (écrits en italique) ont une mauvaise corrélation, puisqu'aucune fréquence n'apparaît dans MC monitor,
- 13 MC (écrits normalement) se situent entre les deux cas explicités ci-dessus. Il faudrait donc les étudier au cas par cas afin de les associer à l'un ou l'autre.

Cette analyse réalisée pour trente lignes est également applicable au tableau dans sa totalité.

Maintenant, observons les résultats à l'échelle du tableau en entier (Tab. 7). Les occurrences égales à 0 et 1 représentent respectivement 17% et 10% des deux listes, ce qui est un pourcentage important.

	MC index		MC monitor	
Occurrence (x)	**Nombre de MC**	**Pourcentage**	**Nombre de MC**	**Pourcentage**
x<=1	381	17%	13	10%
1<x<10	553	24%	76	60%
10<=x<100	714	31%	36	29%
100<=x<1000	479	21%	1	1%
1000<=x	167	7%	0	0%
Total	2294	100%	126	100%

Tab. 7

Dans le tableau suivant (Tab. 8), on peut voir 28 termes ayant une fréquence supérieure ou égale à 10 dans MC monitor. On peut ainsi observer leur rang dans MC index.

Index		Mots-Clés	Monitor	
occurence	rang		rang	occurence
18579	**4**	INTERPRETATION	**5**	*49*
10109	**11**	PRESSION	**7**	*34*
8466	**13**	FORAGE	**11**	*24*
8269	**15**	SISMIQUE	**1**	*102*
8179	**16**	DIAGRAPHIE	**4**	*53*
7039	**21**	TECTONIQUE	**16**	*19*
6705	**23**	CORRELATION	**30**	*12*
6315	**25**	SEDIMENTOLOGIE	**24**	*15*
5671	**33**	SISMOSONDAGE	**37**	*11*
4654	**43**	GEOLOGIE	**8**	*32*
4644	**44**	COMPLETION	**13**	*21*
3577	**62**	GEOPHYSIQUE	**9**	*29*
2234	**91**	PRODUCTION	**15**	*21*
2079	**99**	SISMIQUE 3D	**20**	*16*
1795	**109**	CAROTTAGE	**27**	*12*
1720	**114**	EAU	**31**	*12*
1206	**153**	TRAITEMENT	**22**	*16*
1178	**155**	HISTORIQUE	**34**	*12*
848	**183**	PUITS	**2**	*96*
681	**214**	FAISABILITE	**32**	*12*
601	**233**	GAZ	**33**	*12*
454	**285**	EXPLOITATION	**38**	*10*
351	**337**	GRABEN	**12**	*24*
283	**379**	DELTA	**18**	*16*
272	**385**	ASSOCIATION	**26**	*12*
135	**551**	DESHUILAGE	**19**	*16*
31	**990**	CONFERENCE	**29**	*12*
8	**1396**	CAPITAL	**23**	*15*

Tab. 8

(Dans ce tableau, il manque certains numéros de rang dans MC monitor. Cela est dû à un problème de synonymie. En effet, dans MC index, les relations de synonymie ont été enlevées, par contre, les interrogateurs continuent de les utiliser.)

b- Application des deux méthodes (cf. III)

Dans la partie III, nous avions détaillée deux méthodes : l'une « pragmatique » et l'autre «réfléchie ». Mettons les maintenant en œuvre.

- Pour les MC ayant une fréquence de 0 ou 1, il est clair que leur élimination pure et simple serait un début probant pour une révision du thesaurus. En effet, ils représentent respectivement 17% et 10% des deux listes (cf. Tab. 7). La méthode « pragmatique » serait donc particulièrement efficace.

- Pour les MC ayant une fréquence de 1 à 10, le choix est plus difficile. Si l'on applique la méthode « pragmatique » sur les MC index, on peut encore éliminer 24%. Plus de 50% du thesaurus général subsisteraient. Par contre, le fait que l'un de ces mots puisse être un concept important pour l'EP n'est aucunement pris en compte.

- Si l'on prend en compte, la méthode « réfléchie », un fait est clairement mis en évidence : peu de MC sont utilisés lors des interrogations par rapport au nombre de MC présents dans l'index. Sinon cette méthode permet une approche plus fine. La figure 20 permet d'observer qu'un terme très interrogé n'est pas toujours le plus indexé (ex : deshuilage : 19ème rang dans MC monitor et 551ème dans MC index).

Ces résultats doivent être relativisés car les deux listes ont une différence importante de termes. Néammoins, ni la corrélation ni la non-corrélation des deux listes n'est flagrante. Il semble donc possible de déclarer que la méthode « pragmatique » est trop « expéditive » et que la méthode « réfléchie » apporte matières à réflexion..

8- Conclusion de l'Etude

D'après les résultats obtenus et les observations réalisées, on peut donc conclure que **la pertinence d'un terme n'est pas *toujours* liée à sa fréquence**. La méthode « pragmatique » doit donc être utilisée avec parcimonie. Par contre, il est certain que le **thesaurus de l'EP a besoin d'être révisé.** Ainsi, les mots-clés indexés et interrogés seront mieux corrélés dans le futur. Pour l'instant, la solution proposée est la suivante :

- les fréquences d'apparition égales à 0 doivent disparaître,
- les fréquences d'apparition égales à 1 impliquent la transformation de leur MC en ML (mot libre),
- les MC, ayant une fréquences d'apparition bien corrélée, sont conservées,
- les autres doivent être étudiés cas par cas, ou en tout cas, groupe par groupe, pour savoir s'ils sont conservés ou transformés en ML.

De plus, il faut tenir compte qu'une mauvaise corrélation peut refléter les deux situations suivantes :

- le terme est très spécifique du domaine, mais il n'apporte aucun renseignement pour une recherche,

  Ex : pétrole brut doit être présent dans le domaine de l'EP car énormément de rapports vont traiter de ce sujet, mais il sera peu pertinent lors d'une interrogation

- notre choix de base a été de 6 mois, il faudrait peut-être réaliser cette étude sur une période plus longue avant de réellement parler de termes superflus, s'ils ne sont pas interrogés.

Par conséquent, la méthode **pour réviser un thesaurus** provient d'une **réflexion** qui repose sur **plusieurs propositions**.

Dans la cadre du projet Infodep, lorsqu'une méthodologie sera clairement définie, il sera intéressant de l'appliquer au thesaurus d'Infodop. Ensuite, il faudra étudier Infodex afin d'établir les tables de correspondance. Les thesaurii seront ainsi révisés et pourront être migrés vers Infodep.

Le thesaurus d'Infodep sera spécifique à l'EP, mais la base en elle-même contiendra toujours autant de références. En effet, les MC éliminés le seront soit définitivement, soit ils seront transformés en mots libres. Le projet ensuite continuera sur l'étude des mots libres. En effet, les plus fréquemment utilisés seront peut-être transformés en mots-clés contrôlés, mais ceci est une autre histoire...

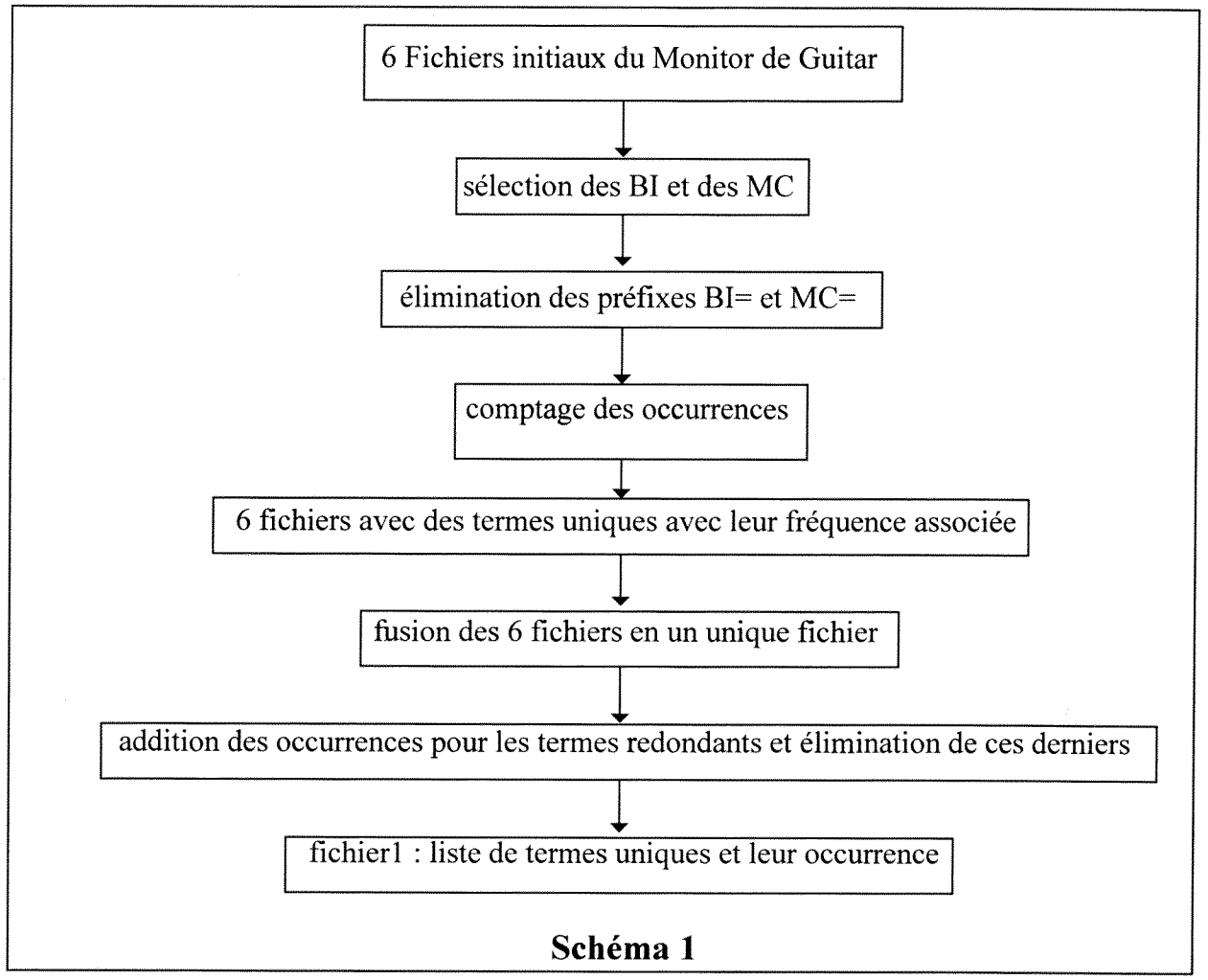


Schéma 1

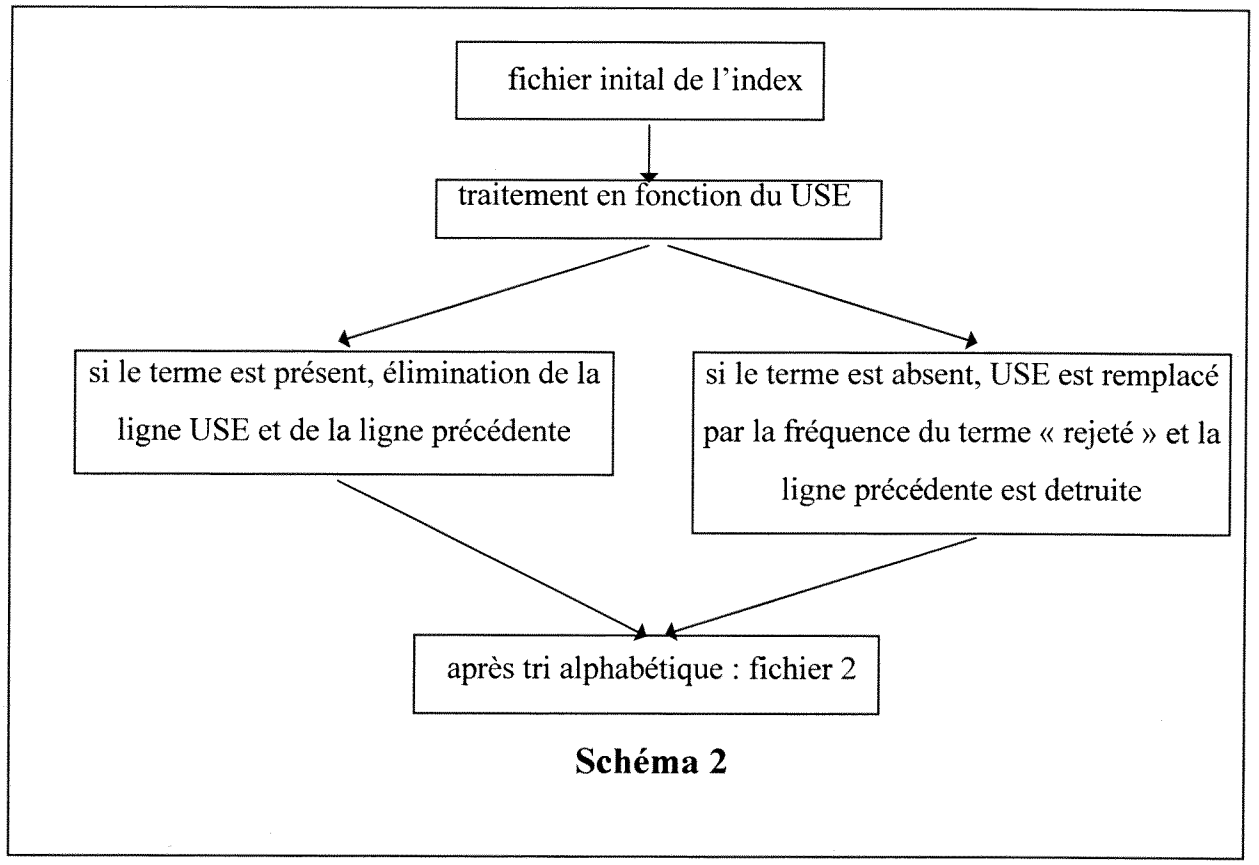


Schéma 2

L'opération suivante sera réalisée avec fichier1 et fichier2.

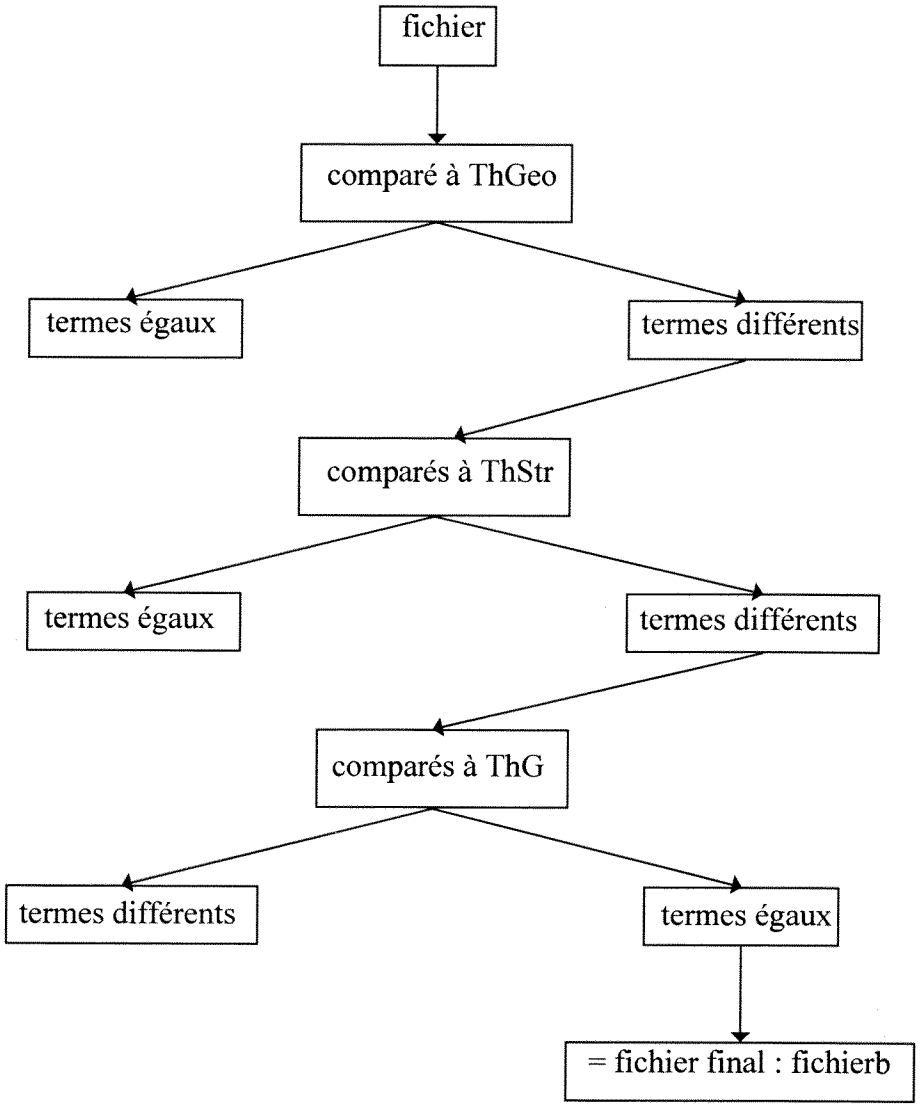


Schéma 3

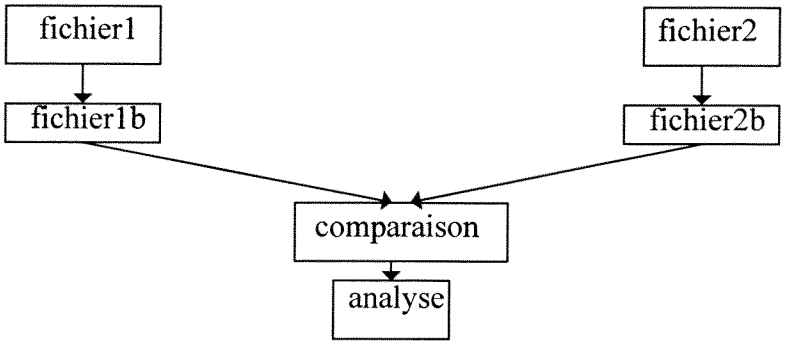


Schéma 4

V- CONCLUSION

⇨ La révision du thésaurus de la base de données Guitar s'inscrit dans un projet plus large : la mise en place de l'Outil de Manipulation des Données (OMD). Grâce à lui, l'utilisateur final aura accès à de nombreuses bases à travers une interface unique. Mais, lorsque l'on parle d'un projet dans sa globalité, on n'a pas toujours conscience de ses composantes. Ainsi le but de mon stage, d'une certaine manière, aura donc été de mieux révéler l'une d'entre elles : la migration-fusion de plusieurs bases de données. Il m'a fallu découvrir quelle méthode employer pour simplifier un thesaurus. La démarche réalisée au cours de ce stage est un regard subjectif porté sur ce problème. Néanmoins, elle n'est pas et n'a jamais voulue être « la » solution. En effet, si l'on désire obtenir un résultat constructif lors de la révision d'un thesaurus, il n'y a pas de méthode radicale.

⇨ Dans ce mémoire, deux méthodes vous ont été proposées. Elles ne sont ni meilleures ni moins bonnes que d'autres. Mais j'ai trouvé passionnant de les construire et de les appliquer pour arriver au tableau final de comparaison. Celui-ci constitue un résultat à exploiter : il suggère des pistes tangibles. En effet, pour s'engager dans un processus, il est préférable d'avoir plusieurs propositions sur lesquelles s'appuyer afin d'agir au mieux. Cependant, une certitude se dégage de cette étude : la nécessité d'un thesaurus plus spécifique à l'Exploration-Production. Ainsi, dans le futur, arrivera-t-on à des mots-clés aussi pertinents en indexation qu'en interrogation car ils cerneront vraiment les domaines de l'EP.

⇨ A un niveau plus personnel, ce stage m'a beaucoup appris. En effet, apporter ma pierre à un édifice aussi ambitieux que cette migration-fusion de bases de données internes a été très stimulant. Ce projet m'a également permis de mesurer toute l'utilité, dans un environnement documentaire scientifique, de ma double compétence (géologie et documentation), même si au cours de ce stage j'ai surtout mis en pratique et affiné mes connaissances en documentation.

⇨ Une étape a ainsi été franchie, mais un travail de longue haleine se poursuit pour les personnes responsables de l'OMD.

BIBLIOGRAPHIE

AFNOR et al. *Gérer et assurer la qualité.* Paris. AFNOR, 1986. 770p.

DUVERGE, J.M. *Base de données INFODEX - THESAURUS.* Pau. Elf Exploration Production, 1995. 250p.

JURAN, J. *Gestion de la Qualité.* AFNOR, 1983. 517p.

MICHAUT, B. et al. *Introduction à l'exploration pétrolière.* ENSPM, 1996. 320p.

MILLS, C.A. *The Quality Audit : a management evaluation tool.* New-York. McGraw-Hill Publishing, 1989. 305p.

MOUVEMENT FRANCAIS DE LA QUALITE. *Synthèse du groupe de travail indicateur qualité et tableau de bord.* Mouvement Français de la Qualité, 1994. 136p.

REED J. *Etape par étape Microsoft Excel 97 Visual Basic.* Microsoft Press, 1997. 329p.

SADO, G. et SADO, M.C. *Les plans d'expériences de l'expérimentation à l'assurance qualité.* Institut d'études et de développement, 1991. 13p.

SIGLER L. *Guitar/Libra/Gazet : Manuel d'utilisation.* Société Nationale Elf Aquitaine (Production), 1991. 59p.

SIGLER, L. *Initiation à la gestion documentaire dans l'entreprise.* Elf Aquitaine Production, 1995. 71p.

THESAURUS IDELFA. Elf Aquitaine, 1991. 7ème édition.