N° d'ordre : 244

THÈSE

présentée à

l'UNIVERSITÉ DES SCIENCES ET TECHNIQUES de LILLE I

pour obtenir le titre de

DOCTEUR DE SPÉCIALITÉ

(Mathématiques appliquées)

par

Thierry GIRARD

Conception et expérimentation d'une banque de données économiques

UNIVERSITÉ DES SC. ET TECHN. BIBLIOTHÈQUE UNIVERSITAIRE LILLE

THÉSE soutenue le 30 Avril 1971,

devant la Commission d'examen :

Monsieur P. BACCHUS, Président

Monsieur B. DRIEUX, examinateur

Monsieur J.-P. TRYSTRAM, examinateur

Monsieur C. CARREZ, rapporteur

UNIVERSITE DE LILLE

FACULTE DES SCIENCES

DOYENS HONORAIRES

MM. H. LEFEBVRE, M. PARREAU.

PROFESSEURS HONORAIRES

MM. ARNOULT, BROCHARD, CAU, CHAPPELON, CHAUDRON, CORDONNIER, DEHEUVELS, DEHORNE, DOLLE, FLEURY, P. GERMAIN, KAMPE DE FERIET, KOURGANOFF, LAMOTTE, LELONG, Mme LELONG, MM. MAZET, MICHEL, NORMANT, PARISELLE, PAUTHENIER, ROIG, ROSEAU, ROUBINE, ROUELLE, WIEMAN, ZAMANSKY.

PROFESSEURS TITULAIRES

M. BACCHUS Pierre	Astronomie et Calcul
M. BEAUFILS Jean-Pierre	Chimie Générale
M. BLOCH Vincent	Psychophysiologie
M. BONNEMAN Pierre	Chimie Industrielle
M. BONTE Antoine	Géologie Appliquée
M. BOUGHON Pierre	Mathématiques
M. BOURIQUET Robert	Biologie Végétale
M. CELET Paul	Géologie Générale
M. CONSTANT Eugène	Electronique
M. CORSIN Pierre	Paléobotanique
M. DECUYPER Marcel	Mathématiques
M. DEDECKER Paul	Mathématiques
M. le Doyen DEFRETIN René	Directeur du Laboratoire de Biologie Maritime de Wimereux
M. DELATTRE Charles	Géologie Générale
M. DURCHON Maurice	Biologie Animale
M. FOURET René	Physique
M. GABILLARD Robert	Electronique
M. GLACET Charles	Chimie Organique
M. GONTIER Gérard	Mécanique des Fluides
M. GUILLAUME Jean	Biologie Végétale
M. HEUBEL Joseph	Chimie Minérale
Mme LENOBLE Jacqueline	Physique
M. MONTREUIL Jean	Chimie Biologique
Mme SCHWARTZ Marie Hélène	Mathématiques
M. TILLIEU Jacques	Physique
M. TRIDOT Gabriel	Chimie Minérale Appliquée E.N.S.C.L.
M. VIDAL Pierre	Automatique
M. VIVIER Emile	Biologie Animale
M. WATERLOT Gérard	Géologie et Minéralogie
M. WERTHEIMER Raymond	Physique

PROFESSEURS A TITRE PERSONNEL

M. BOUISSET Simon	Physiologie Animale
M. DELHAYE Michel	Chimie Physique et Minérale 1er Cycle
M. LINDER Robert	Biologie Végétale
M. LUCQUIN Michel	Chimie Physique
M. PARREAU Michel	Mathématiques
M. SAVARD Jean	Chimie Générale
M. SCHALLER François	Biologie Animale
M. SCHILTZ René	Physique

PROFESSEURS SANS CHAIRE

M. BELLET Jean	Physique
M. BODART Marcel	Biologie Végétale
M. BOILLET Pierre	Physique
M. DERCOURT Jean-Michel	Géologie et Minéralogie
M. DEVRAINNE Pierre	Chimie Minérale
Mle MARQUET Simone	Mathématiques
M. MONTARIOL Frédéric	Chimie Minérale Appliquée
M. PROUVOST Jean	Géologie et Minéralogie
M. VAILLANT Jean	Mathématiques

MAITRES DE CONFERENCE (et chargés des fonctions)

M. AUBIN Thierry	Mathématiques Pures
M. BEGUIN Paul	Mécanique des Fluides
M. BILLARD Jean	Physique
M. BKOUCHE Rudolphe	Mathématiques
M. BOILLY Bénoni	Biologie Animale
M. BONNOT Ernest	Biologie Végétale
M. CAPURON Alfred	Biologie Animale
M. CARREZ Christian	Calcul
M. CORDONNIER Vincent	Calcul
M. CORTOIS Jean	Physique
M. COULON Jean-Paul	Electrotechnique
M. DEBRABAN Pierre	Sciences Appliquées
M. ESCAIG Bertrand	Physique
M. FROLICH Daniel	Sciences Appliquées
M. GOBLOT Rémi	Mathématiques
M. GOUDMAND Pierre	Chimie Physique
M. GRUSON Laurent	Mathématiques
M. GUILBAULT Pierre	Physiologie Animale
M. HERMAN Maurice	Physique
M. HUARD de la MARRE Pierre	Calcul
M. JOURNEL Gérard	Sciences Appliquées
Mle KOSMANN Yvette	Mathématiques
M. LABLACHE COMBIER Alain	Chimie Générale
M. LACOSTE Louis	Biologie Végétale
M. LANDAIS Jean	Chimie Organique
M. LAURENT François	Automatique
M. LEHMANN Daniel	Mathématiques
Mme LEHMANN Josiane	Mathématiques
M. LOCQUENEUX Robert	Physique
M. LOUAGE Francis	Sciences Appliquées

M. LOUCHEUX Claude	Chimie Physique
M. MAES Serge	Physique
M. MAIZIERES Christian	Automatique
M. MESSELYN Jean	Physique
M. MIGEON Michel	Sciences Appliquées
M. MONTEL Marc	Physique
M. OUZIAUX Roger	Sciences Appliquées
M. PANET Marius	Electrotechnique
M. PAQUET Jacques	Sciences Appliquées
M. PARSY Fernand	Mécanique des Fluides
M. POVY Jean-Claude	Sciences Appliquées
M. RACZY	Radioélectrique
M. ROUSSEAU Jean-Paul	Biologie Animale
M. ROYNETTE Bernard	Mathématiques
M. SALMER Georges	Electronique
M. SMET Pierre	Physique
M. VANDORPE Bernard	Sciences Appliquées
M. WATERLOT Michel	Géologie Générale
Mme ZINN **JUSTIN** Nicole	Mathématiques

Je tiens à remercier bien vivement le Professeur P. BACCHUS pour avoir accepté de bien vouloir présider ce travail, et mis à ma disposition le Laboratoire de Calcul de la Faculté des Sciences de LILLE.

Pour les conseils qu'ils m'ont constamment prodigués, pour les encouragements de tous les instants qu'ils m'ont apportés, les Professeurs C. CARREZ et B. DRIEUX sont assurés de toute ma reconnaissance.

Que l'on me permette d'exprimer plus particulièrement ma gratitude au Professeur J.P. TRYSTRAM pour sa bienveillante attention à l'égard de mon travail, pour l'expérience et la haute autorité dont il a bien voulu me faire profiter dans le cadre des travaux qu'il a dirigés successivement à la Faculté des Lettres et Sciences Humaines de Lille, et à l'Institut de recherche d'Informatique et Automatique, pour le compte de la Délégation à l'Aménagement du Territoire et à l'Action Régionale.

S O M M A I R E

INTRODUCTION

PARTIE I : NOTIONS GENERALES

Chapitre 1 - Bases de données
Chapitre 2 - Banques de données
Chapitre 3 - Sciences humaines et banques de données

PARTIE II : LE SYSTEME D'ACCUMULATION DES DONNEES

Chapitre 1 - Le système S.A.D. I
Chapitre 2 - Le système S.A.D. II
Chapitre 3 - Le système S.A.D. II T.S.

CONCLUSION

Je tiens à remercier Marie-France RABOURDIN et Michèle DRIESSENS dont la gentillesse et le dévouement ont permis la réalisation matérielle de ce document.

INTRODUCTION

La rencontre de l'informatique et des sciences humaines résulte en bonne part de l'évolution de l'informatique quant à ses techniques, ses moyens matériels et de l'évolution générale des sciences sociales au point de vue de la diffusion des connaissances, de l'abondance des sujets abordés, des méthodes mises en oeuvre. Il faut d'ailleurs noter que l'évolution des méthodes quantitatives des sciences humaines n'aurait pu, sans doute, avoir lieu sans l'apparition de l'informatique. L'effet induit par l'utilisation de celle-ci est cumulatif.

Le premier problème posé par les sciences humaines est soulevé par l'extrême diffusion des connaissances qu'ont connue ces disciplines. Le chercheur ne peut facilement assurer son information parmi toute la documentation disponible. La vie sociale et économique est toujours une source d'information extrêmement féconde. L'étude généralisée des objets historiques, surtout dans le domaine quantitatif, a mis en évidence l'énorme quantité d'informations à traiter par les chercheurs.

Les progrès enregistrés dans les domaines d'application des sciences humaines ont été également générateurs d'informations. La généralisation des enquêtes, sondages, élections, a fait croître les besoins de traitement d'une information plus facile à saisir aujourd'hui qu'autrefois.

Enfin, l'extension des méthodes quantitatives introduites dans les sciences humaines a engendré un besoin considérable en matière de traitement de l'information. L'utilisation de "modèles", fondée sur l'usage des statistiques, est généralisée à beaucoup de disciplines des sciences humaines.

L'informatique a maintenant débordé le domaine du calcul scientifique et technique et le domaine de la simple gestion. Elle apporte son concours fructueux aux spécialistes des sciences humaines essentiellement en raison de trois facteurs :

- les moyens d'archivages ont été considérablement augmentés. La technologie des mémoires de masse évolue de telle façon que le coût de stockage diminue par rapport à la quantité d'information stockée. D'autre part, les techniques d'accès aux informations ont permis d'obtenir des temps de recherche très faibles, en partie indépendants du volume de l'information enregistrée.

- Les techniques d'utilisation des ordinateurs et la naissance de nouveaux langages permettent à l'utilisateur non spécialiste de l'information d'utiliser une grande variété de ressources en matière d'accès et de traitement.

Il est maintenant possible, au prix d'une formation légère, pour le spécialiste en sciences humaines, d'utiliser sans intermédiaire de nombreux systèmes et langages, lui permettant de prendre connaissance de l'information disponible et de lui faire subir, le cas échéant, un certain nombre de traitements. Ceci est facilité par l'apparition de systèmes fonctionnant en "temps réel". Les techniques conversationnelles ont fait perdre à l'usage de l'ordinateur une rigidité qui faisait écran.

D'autre part, les possibilités du télétraitement sont un facteur de diffusion de la connaissance en mettant à la portée de tous des moyens allégeant considérablement les tâches du chercheur. Celui-ci, disposant de divers matériels lui permettant de diversifier les représentations de l'information et de "dialoguer" plus aisément avec l'ordinateur, peut plus

facilement s'affranchir de tâches qui, accessoires dans son activité, n'en étaient pas moins lourdes et génératrices d'erreurs.

Vis-à-vis des disciplines relevant des sciences humaines, l'informatique a principalement deux fonctions. La première est une fonction d'archivage. Cette fonction est liée aux possibilités de sélection de l'information ; c'est la fonction documentaire qui peut s'excercer sur des "objets divers, tels que documents, fichiers, tableaux statistiques. La deuxième est la fonction de traitement. Ces traitements sont de tous ordres, depuis la simple tabulation, jusqu'à l'exploitation de fichiers par des "modèles" très sophistiqués. En matière de recherche d'information disponible dans l'information archivée, l'utilisateur doit disposer de moyens de recherches précises, "tous azimuths", exhaustives autant que faire se peut. En ce qui concerne les traitements, l'utilisateur en sciences humaines a des exigences de fiabilité de l'information et de souplesse dans la mise en oeuvre des traitements. Il faut noter également que le faible coût marginal des traitements sur un ensemble de données permet de rentabiliser, dans certains cas, l'investissement fait dans la saisie des données.

L'utilisation de l'informatique par les sciences humaines est requise par le très grand volume des informations manipulées par les spécialistes de ces disciplines et par la croissance de leurs besoins en matière de traitements mathématiques et statistiques. En vue de faire partager l'information disponible et les moyens de traitements, pour diminuer les coûts de stockage et de temps de calcul, pour assurer pleinement les fonctions d'archivages que les moyens classiques ne pourront bientôt plus remplir, on a commencé à construire des banques de données. Ces systèmes, dont la première génération a vu le jour dans le début des années 60, se généralisent. Leurs créateurs ont des objectifs très divers. Certains systèmes sont conçus pour permettre l'archivage et le traitement d'informations d'ordre historique, politique, économique ou social. L'expérience de l'une de ces banques de données sera décrite après un chapitre qui constitue un rappel

des techniques mises en oeuvre dans la création, la gestion et l'utilisation de telles banques de données. Les banques de données relatives aux sciences humaines posent un certain nombre de problèmes particuliers du fait de la nature de l'information et de son volume. Ces problèmes seront évoqués avant la description du système SAD.

Partie I

NOTIONS GENERALES

Chapitre I

BASES DE DONNEES

1.1. - Notion de fichier

On appelle fichier toute collection finie d'informations. Cette collection d'informations est généralement munie d'une structure et elle est enregistrée sur un support. Il peut exister deux informations particulières dans le fichier :

a) une information caractérisant le fichier et permettant de le localiser sur son support, d'y accéder et de définir son mode d'enregistrement et son encombrement,

b) une information permettant de la borner sur son support.

Un fichier est enregistré sur un support et cela peut être fait de plusieurs manières. Les supports peuvent être classés de différents points de vue. Tout d'abord, le support peut être réutilisable ou non réutilisable. Un support peut être séquentiel ou adressable. Il est séquentiel si, pour accéder à une information, il est obligatoire de lire toutes les informations qui la précèdent sur le support. Au contraire, il est adressable si l'information est accessible par un adressage (connaissance d'une adresse).

Un fichier est une collection d'enregistrements. Un enregistrement est l'ensemble d'informations dont le nombre est le plus petit qui puisse être transféré entre la mémoire centrale et le support. Dans le cas d'un fichier sur liste d'imprimante, un enregistrement est une ligne, dans le cas d'un fichier carte, une carte, etc ...

Un bloc est composé d'un ensemble d'enregistrements dont le nombre dépend du support. Pour les supports séquentiels le bloc contient un

enregistrement. Pour les supports adressables un bloc est composé de un ou plusieurs enregistrements. Un bloc peut contenir un ou plusieurs articles.

On appellera élément toute information élémentaire du fichier, donc non séparable en éléments. On appellera nom toute désignation d'un élément. Une occurrence particulière d'un élément donne une valeur à celui-ci. On appellera classe de valeurs d'un élément l'ensemble des occurrences d'un élément. Les occurrences d'un élément prennent leurs valeurs dans un vocabulaire et peuvent être de longueur fixe ou variable.

On appellera groupe un ensemble défini d'éléments. On appellera groupe répétitif un groupe qui peut figurer plusieurs fois. On appellera groupe hiérarchisé un ensemble défini de groupes.

On appellera article un ensemble de groupes hiérarchisés ou non, et fichier un ensemble d'articles.

L'information présentée par un élément peut être une information de base du fichier, ayant un contenu sémantique propre dans la structure sémantique du fichier. Elle peut aussi être une donnée sur un autre élément. Elle prend alors le nom de référence ou pointeur.

Ces références peuvent se classer en trois catégories :

- références internes (à un article),
- références externes inter-articles,
- références inter-fichiers.

Ces références définissent des structures sémantiques des fichiers. Elles sont utilisées pour la consultation, la mise à jour et sont construites lors de la création du fichier. Les deux dernières sont importantes car elles

permettent la structuration des fichiers et bases de données.

1.2. - Méthodes de Rangement et Accès.

L'utilisation des fichiers se ramène toujours à une recherche d'article pour accéder à l'information qu'il représente en partie ou en totalité. La recherche dépend du rangement des articles et l'accès s'en trouve profondément influencé.

On divise les rangements en deux classes :

- le rangement séquentiel : les articles sont placés sur leur support les uns après les autres. Ils sont en général classés par référence à la valeur d'un élément , d'un groupe, appelée clé.

- Le rangement permettant un accès direct : l'accès est direct par rapport à un élément si on peut accéder à l'ensemble des articles qui contiennent l'élément, pour une occurrence donnée de celui-ci, ou, plus généralement, qui contiennent l'élément tel qu'il satisfasse à une relation donnée.

Les accès alors peuvent se répartir en 3 classes :

- l'accès séquentiel :

L'accès est séquentiel lorsque le rangement est séquentiel et lorsqu'on procède à la lecture de tous les articles qui précèdent l'article recherché.

- L'accès direct :

L'accès est direct si, connaissant l'occurrence d'un élément, on peut déterminer l'adresse des articles qui contiennent l'élément satisfaisant à la condition imposée.

- L'accès séquentiel indexé :

On peut ranger les articles par classe d'équivalence de valeur de clé. L'accès à une classe est direct. On accède alors à l'article de la classe atteinte par consultation séquentielle. La relation entre la valeur de la clé et l'adresse de la classe d'équivalence est généralement donnée par une table d'index. Cette table constitue un fichier dont l'article est à 2 éléments : la clé, et un pointeur. Cette table peut également être organisée comme un fichier à accès séquentiel indexé. La table d'index de ce fichier est alors souvent appelée "micro-table". L'accès est alors à 2 niveaux. Un fichier à accès séquentiel-indexé inclut donc, outre l'information propre, les tables d'index nécessaires à sa consultation. La recherche d'articles sur plusieurs clés nécessite la présence de plusieurs tables d'index correspondant à chacune des clés.

1.3. - Structures

Les structures décrites sont introduites par l'utilisation des fichiers. Autrement dit, à côté de structures syntaxiques précédemment décrites, il est possible de caractériser les fichiers par des structures sémantiques. Ces structures ne sont pas indépendantes de la méthode de rangement des articles et du mode d'accès.

- Le chaînage.

Cette méthode consiste à ajouter dans l'article des éléments ou groupes de références (pointeurs) qui pointent, en donnant une adresse, vers un autre article qui est équivalent ou précédent du point de vue d'une relation entre clés.

Une méthode souvent employée d'accès direct à un fichier consiste

à calculer l'adresse d'un article d'après la valeur d'un élément de nature numérique. Si ce n'est pas le cas, on construit par diverses méthodes un élément numérique qui lui correspond (méthodes numériques, hash - coding). Ceci constitue la clé. On peut alors calculer l'adresse, en supposant les enregistrements de longueur fixe, par la méthode du reste de la division de la clé par le nombre d'enregistrements qu'il est possible de placer dans la zone prévue. Ce reste constitue l'adresse relative de l'enregistrement dans la zone. Il est évident que plusieurs articles différents pour la valeur de la clé donneront un reste identique. On divise la zone prévue en deux parties, une zone principale et une zone de débordement. La zone principale reçoit à leur adresse les enregistrements dans l'ordre d'arrivée. S'il se présente un second enregistrement, on ajoute à l'article enregistré le premier un élément dont la valeur sera l'adresse dans la zone de débordement où l'on rangera l'article en cours. Cette zone de débordement ne sera accessible que par le pointeur d'enregistrement. Au second cas d'égalité d'adresse, le chaînage est assuré par l'article de la zone de débordement vers l'article en cours placé à la première adresse disponible de la zone de débordement. Dans cette zone, le dernier article d'une classe d'adresse principale est également pourvu d'un élément de chaînage dont la valeur est par convention de l'utilisateur significative de la fin de chaîne.

On peut aussi envisager le cas où le calcul sur la clé conduit à l'adresse d'un groupe qui contient les articles de même valeur d'adresse. On peut y ajouter une zone de débordement, adressée par chaînage dans la zone du groupe.

Les applications des chaînages sont multiples, et la structure de ceux-ci peut être complexe. Les difficultés de création et surtout de mise à jour des chaînes peuvent être importantes.

- La structure de liste.

Cette structure permet le chaînage d'articles à l'intérieur d'un même fichier ou le chaînage entre articles et sous-fichiers d'un autre fichier. A un article est associé un pointeur qui est l'expression d'une relation donnée.

La liste est un anneau si l'un des pointeurs de la chaîne est l'adresse du premier article.

- La structure d'arbre.

Cette structure consiste pour des articles à groupes hiérarchisés, à associer à l'occurrence d'un groupe un pointeur vers le sous-groupe de niveau inférieur.

- La structure de liste multiple.

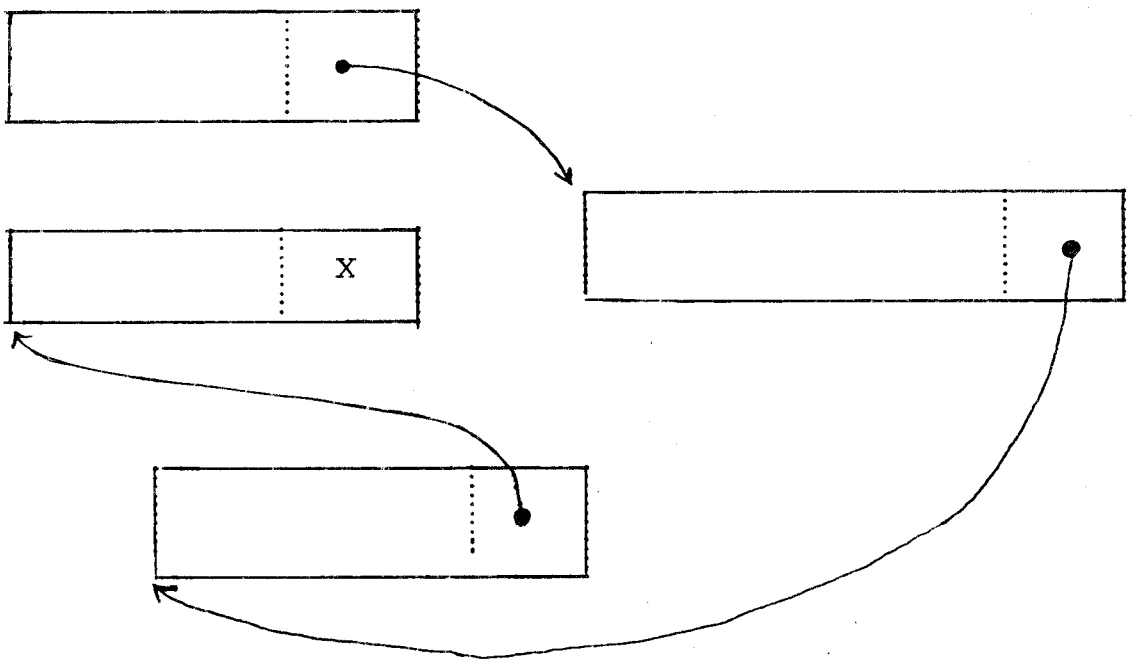
Cette structure dérive de la structure de liste, en multipliant les chaînages entre articles. Chaque chaîne est associée à un groupe de l'article. Un index initialisera les chaînes.

- La structure de liste réciproque.

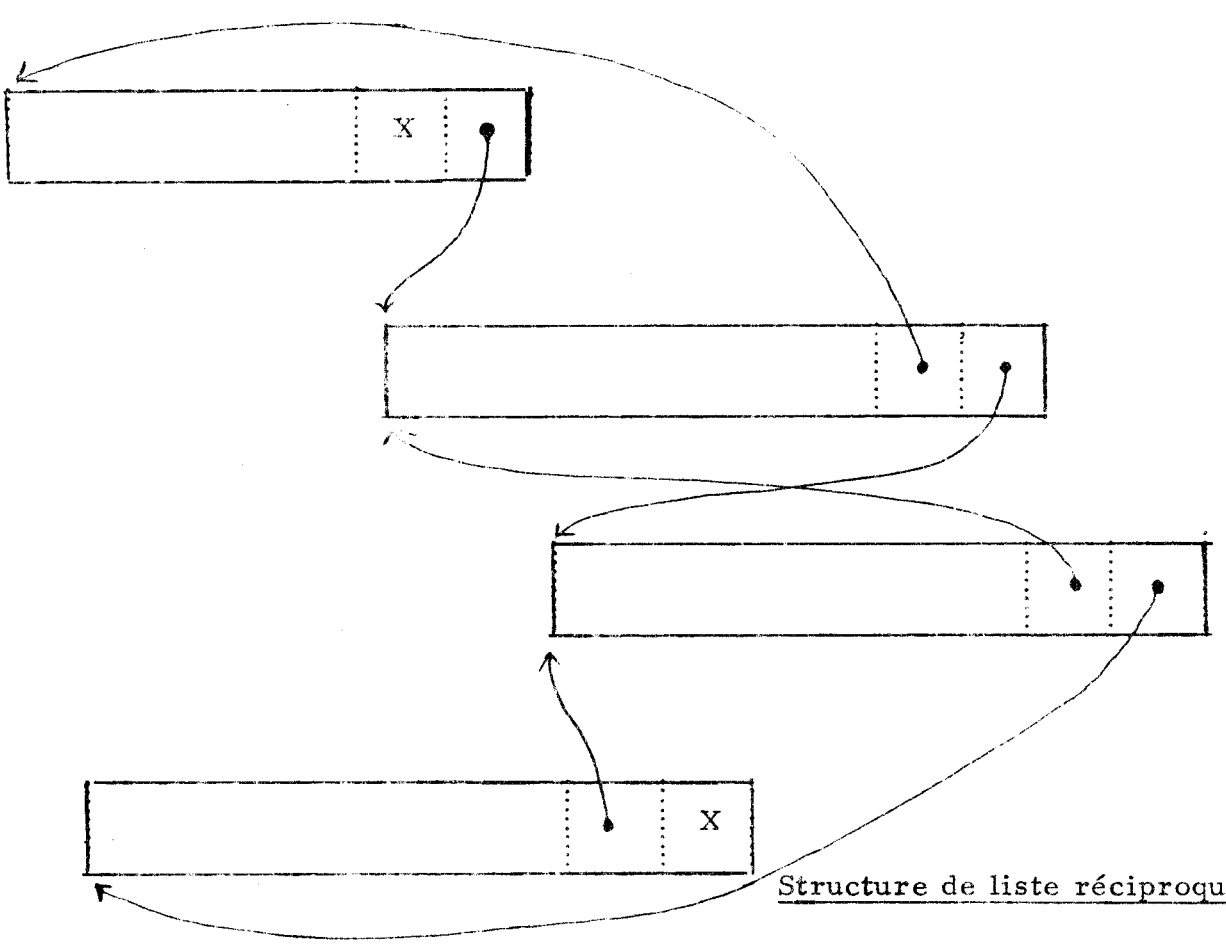
Une liste est dite réciproque si au chaînage de liste on ajoute le chaînage inverse. Pour chaque article, il existe un pointeur vers l'article suivant dans la chaîne, et vers le précédent

- Fichier direct et fichier inversé.

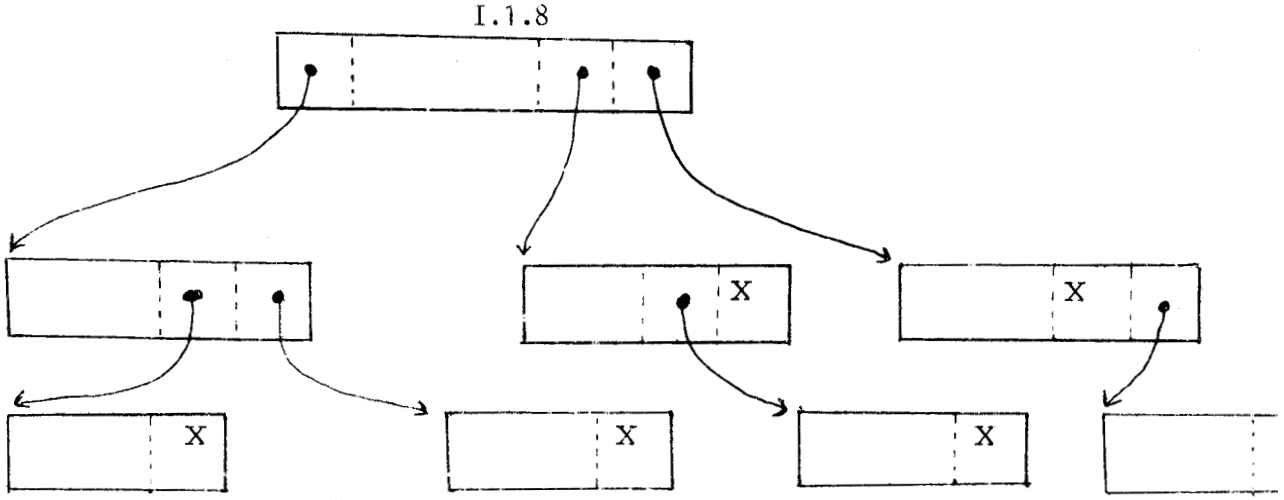
Considérons la structure de liste multiple. La chaîne pour chaque



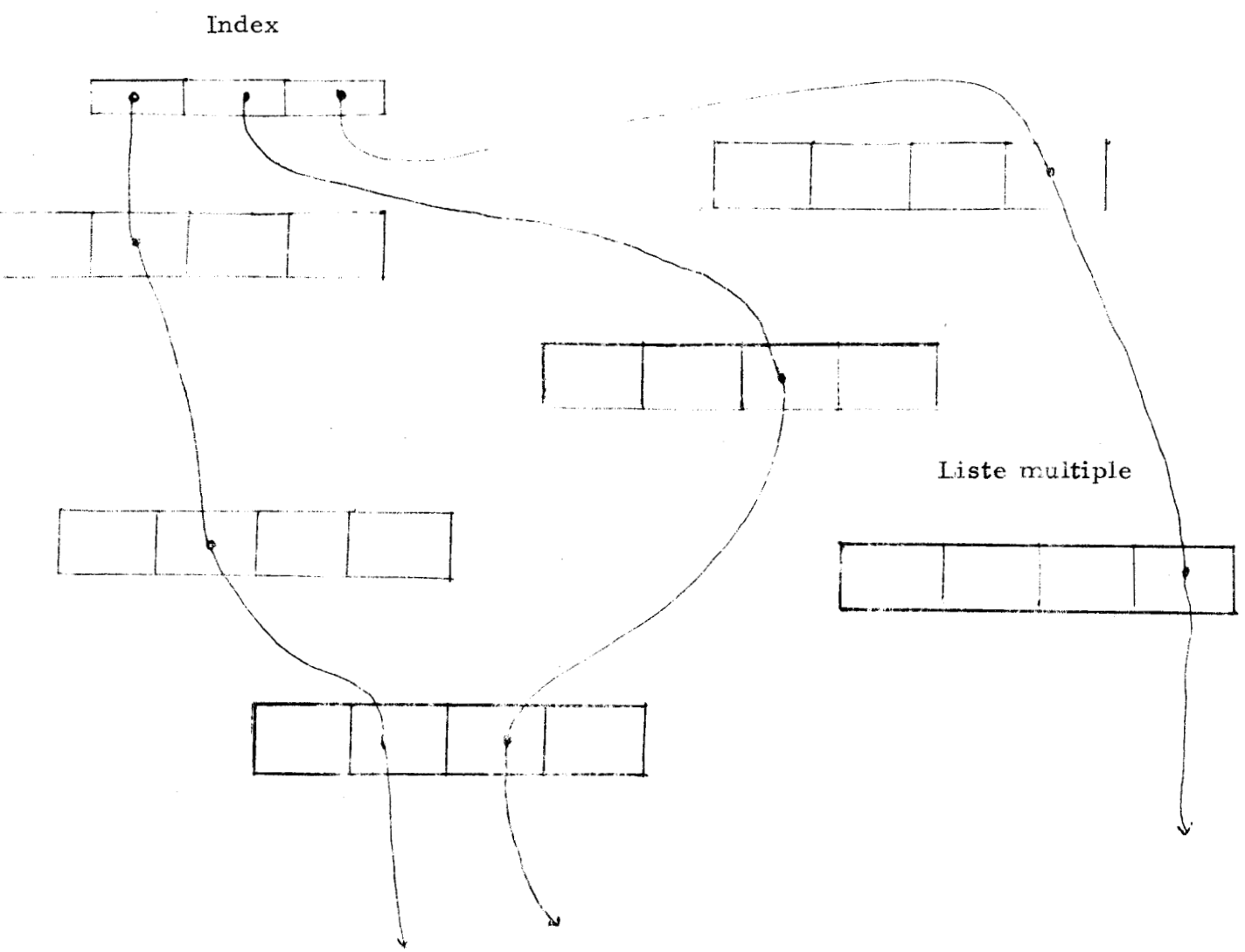
Structure de Liste

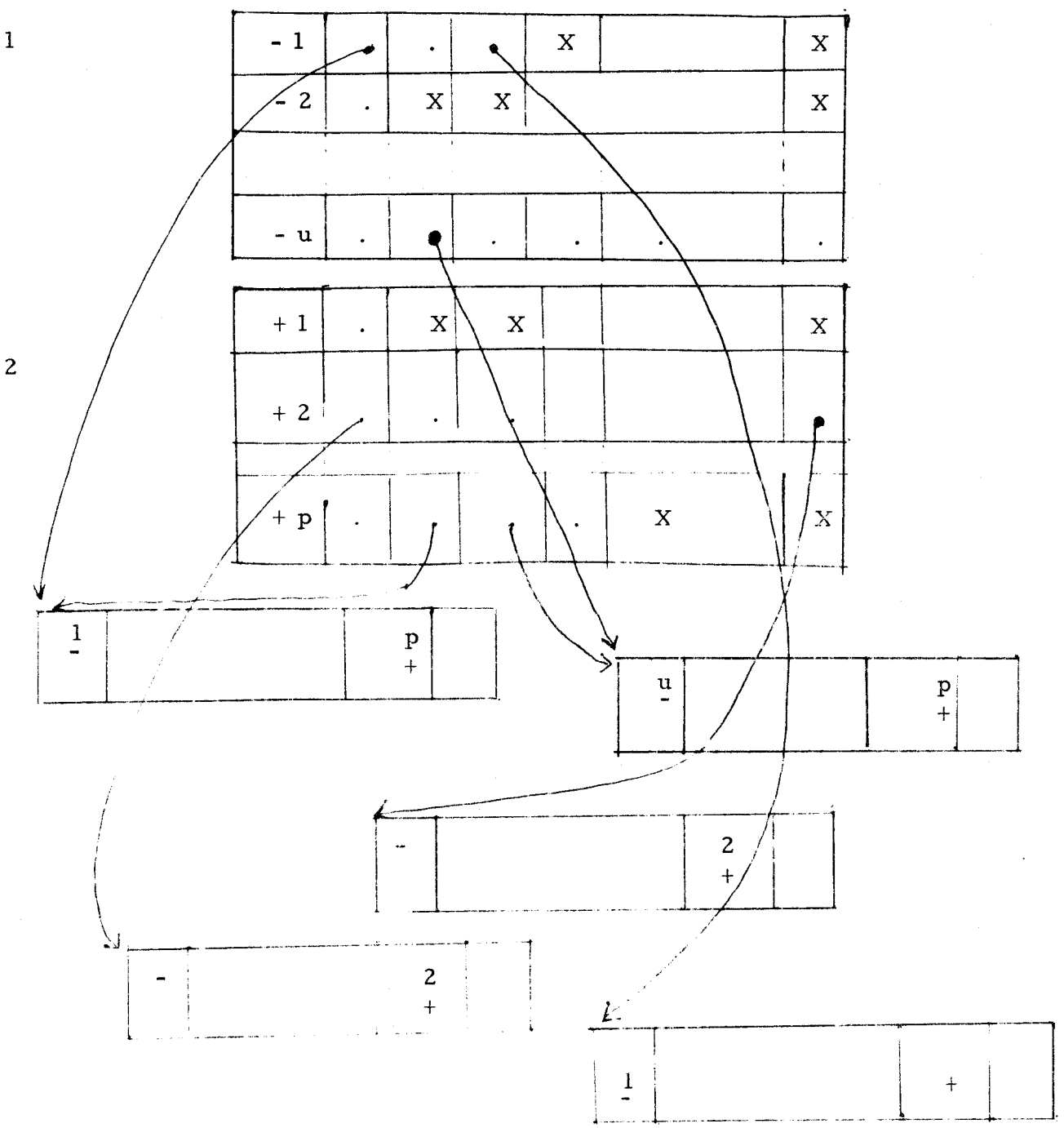


Structure de liste réciproque



Structure d'arbre





Fichier direct et Fichier inversé

occurrence de l'élément ou du groupe est de longueur indéterminée. On peut imposer une longueur en nombre d'articles. .A l'index, on associera les pointeurs des articles initiaux des sous-chaînes. Si l'on considère des sous-chaînes de longueur un, on dit que l'index est inverse. Cet index peut constituer un fichier qui est dit inverse du fichier initial.

1.4. - Caractéristiques des fichiers.

Les fichiers sont caractérisés par :

- le support,
- la méthode de rangement des articles et l'accès,
- la structure logique induite par l'utilisation.

Ces critères sont ceux qui sont définis lors de la création du fichier en vue de son utilisation. Ils influeront sur la structure des programmes de :

- création,
- mise à jour,
- recherche.

Ce sont ces trois actions sur les fichiers qui sont à confronter avec les caractères des fichiers.

1.5. - Base de données

Une base de données est un ensemble de fichiers liés logiquement entre eux afin de permettre la plus large utilisation possible des données qu'ils contiennent. La structure d'une base de données est liée à l'utilisation des fichiers. Cette utilisation peut être multiple et peut conduire à des

structures superposées.

Les fichiers de la base de données se répartissent en 3 classes :

- fichiers de données,
- fichiers de données sur les données,
- fichiers de références.

- Les fichiers de données.

Ce sont les fichiers qui contiennent les informations traitées en vue de l'application Ces fichiers sont organisés de façons diverses et leur utilisation peut nécessiter, pour permettre l'accès aux données, l'existence de références externes.

- Les fichiers de données sur les données.

Pour utiliser des données, il est nécessaire de connaître leur nature syntaxique, sémantique, et leur représentation en mémoire. Ces informations constituent des données qui, du point de vue de l'utilisation des données, peuvent être implicites. Cependant ces informations sont souvent introduites dans la base de données.

C'est, en particulier, le cas de la structure des fichiers. La structure peut être enregistrée également dans chaque article, ce qui permet une plus grande souplesse. Ainsi, la structure d'un fichier peut rester extérieure à un programme utilisateur du fichier.

- Les fichiers de références.

Ces fichiers ont pour rôle d'établir les liens logiques entre fichiers,

ou d'en permettre l'accès.

Dans cette classe de fichiers, on peut définir un ensemble de fichiers particuliers. Ce sont les dictionnaires. Ils assurent la correspondance entre les éléments d'un langage de description de données et un ensemble équivalent du point de vue de l'application.

Enfin, parmi ces fichiers, il est des fichiers qui peuvent se voir assigner des rôles particuliers. Ainsi peut-on trouver des fichiers contenant des informations sur les utilisateurs afin de contrôler l'accès aux fichiers ou même aux articles et groupes.

Les fichiers qui composent la base de données sont gérés par un système. Celui-ci peut nécessiter l'emploi de fichiers particuliers contenant des informations sur la structure actuelle de la base de données, sur la nature des opérations en cours ou passées, par exemple des fréquences d'accès aux données, des statistiques d'utilisations, sur l'état du système au point de vue fiabilité des informations, etc ...

1.6. - Gestion des bases de données

La gestion d'une base de données est celle des fichiers qui composent la base. Elle comporte deux aspects :

- la création,
- la mise à jour.

La gestion peut être assurée, soit par un système de gestion de base de données tout à fait général, soit par un système dépendant étroitement de l'application

- La création.

Il s'agit de créer des fichiers à partir des données, de les organiser. La création d'un fichier est faite selon les systèmes par un programme "batch" ou "on line". Dans le cas d'un programme "batch", celui-ci peut ou non mettre en oeuvre un langage de définition des données. Dans le cas d'existence d'un langage de définition, ce langage assure la structuration des articles et du fichier La méthode de rangement et l'accès seront soit imposées par le système ou laissées au choix de l'utilisatuer si celui-ci a créé le programme en dehors du système de gestion.

Dans le cas d'une création "on line", il y a presque toujours lieu de mettre en oeuvre un langage de définition. Ce langage permet l'écriture de programme dont le rôle est de définir la structure du fichier, de l'article. Il permet aussi la vérification des valeurs, la construction de références conditionnelles et la création d'un vocabulaire des entrées (noms d'articles, de groupes, d'éléments).

- La mise à jour

Elle comprend deux actions distinctes :

- modification de la valeur d'un élément ou d'un groupe dans un article,

- ajout ou suppression d'articles, de groupes ou d'éléments.

Elle nécessite, dans certains cas, l'utilisation de programmes de restructuration de la base de données. La mise à jour peut être faite "on line" ou sur le mode "batch-processing".

Dans le cas "on line", elle se fait par utilisation d'un langage qui

permet d'accéder à l'article et de le modifier. C'est pourquoi, le langage utilisé est souvent celui mis en oeuvre pour l'interrogation et l'exploitation de la base de données

Dans le cas "batch", les programmes de mises à jour utilisent des modules d'accès et de recherches qui peuvent être utilisés par ailleurs, notamment dans des programmes d'applications.

Enfin, l'utilisation de la base de données peut entraîner la nécessité de la restructurer, afin de l'adapter aux besoins des utilisateurs et d'améliorer les performances. Les résultats de certains programmes d'évaluations sont alors exploités pour modifier certains paramètres (tailles de pages, nombre de pages, etc...) ou pour restructurer certains fichiers en fonction de fréquences d'utilisations

La construction d'une base de données est faite en élaborant la structure des fichiers et en choisissant des méthodes d'accès pour chacun d'eux. L'objectif majeur fixé est d'obtenir un temps de réponse acceptable. Le problème de la mise à jour est un critère important pour le choix de la structure des fichiers. Il existe, pour faciliter la solution du problème de la structure de la base de données, des programmes d'évaluations des performances et notamment des programmes de simulation des effets introduits par des structures complexes de fichiers sur l'effacité et les temps de réponses, en matière d'interrogation et de mise à jour. Ces programmes permettent de simuler une configuration de machine et des méthodes d'accès.

1.7. - Etats des fichiers - Protection - Fiabilité.

Tout système de gestion et d'exploitation de base de données doit être envisagé sous l'angle de la sécurité, de la protection de l'information et de sa fiabilité. En effet, ces problèmes se trouvent prendre une grande importance en raison :

- du volume important de données qui peuvent être traitées par ces systèmes,

- de la diversité des utilisateurs,

- de la diversité de nature sémantique, quant à l'aspect des informations traitées.

La sécurité de l'information est liée à la fiabilité du système. Négligeant l'aspect matériel, qui a son importance, on retiendra seulement qu'un système doit pouvoir mettre en oeuvre des procédures exceptionnelles permettant notamment la reconstitution des fichiers de base au cas où ils seraient altérés ou détruits. Un autre cas peut être envisagé, c'est la discordance qui serait introduite par une erreur de programme ou une erreur d'exploitation de ceux-ci. A cet égard, un système qui admet une utilisation partagée de fichier, notamment en mise à jour, doit offrir toutes les garanties de sécurité d'emploi et particulièrement éviter les cas de blocage.

La protection de l'information est liée au secret qui peut marquer certains fichiers ou partie de fichiers d'une base de données. On utilise, pour donner un accès privatif aux données, des techniques de mots de passe à un ou plusieurs niveaux.

La complexité des moyens mis en oeuvre pour atteindre une plus grande sécurité et une meilleure protection dépend évidemment de l'application. Cependant, il faut noter qu'il peut être utile de prévoir, dans un système, l'enregistrement de l'historique des fichiers et de l'état général du système, permettant une meilleure reprise en cas d'incident.

1.8. - Caractérisations d'une base de données.

On caractérisera une base de données par les caractéristiques des fichiers dont elle est formée et par celles du système de gestion.

Les fichiers d'une base de données sont d'abord caractérisés par leur support, mode d'accès et structures. Ils sont aussi caractérisés par leur place dans la structure de la base de données. Ceci peut entraîner, suivant leur utilisation, une diversité de supports et de modes d'accès. La structure de la base de données peut être fondée sur une hiérarchie des rôles de chaque fichier. Cette hiérarchie est souvent fonction de contraintes de temps d'accès et de mémoire utilisable. Ainsi, pour certains fichiers en accès séquentiels indexés, on enregistrera les index sur tambour rapide, et les pages des fichiers sur disques plus lents (EXEC-8). La classe d'application, pour laquelle le système est conçu, peut aussi intervenir dans la définition d'une hiérarchie de l'information.

Le système de gestion se caractérise par :

- le mode d'utilisation des programmes de création et de gestion,

- les techniques de structuration des fichiers de données.

La création et la mise à jour des fichiers peuvent se faire soit "on line", soit en "batch processing".

Le système peut inclure l'utilisation d'un langage de définition de données pour la création. Les performances de ce langage sont examinées du point de vue de la structure d'article créé. Cette structure peut être plus ou moins complexe. Le langage peut autoriser l'inclusion de références multiples et conditionnelles. Un tel langage peut accepter plusieurs type d'informations (chaînes de caractères, numérations diverses). Il peut aussi donner la possibilité de définir des procédures de contrôles à l'introduction des informations, et de placer des clés d'accès et de protection à différents niveaux.

En ce qui concerne la mise à jour, l'utilisateur peut également disposer d'un langage permettant :

- l'accès aux articles,
- la restructuration de ceux-ci,
- la modification d'éléments ou de groupes,
- la modification de références,
- l'ajout d'articles et de références.

Dans un système de fichiers partagés, il peut exister un mécanisme de validation de la mise à jour et un enregistrement des mises à jour dans un fichier historique.

Enfin, on peut distinguer trois classes de bases de données par l'usage qui en est fait :

1 - la base de données est conçue pour une exploitation par un seul système lié étroitement à la base de données.

2 - La base de données est conçue pour permettre l'accès aux données à tout un ensemble de programmes, où le programme d'interrogation est un élément particulier, défini à posteriori.

3 - La base de données est conçue pour la mise en oeuvre d'un système d'exploitation en vue d'applications bien définies ; elle peut être également utilisée pour certains autres types d'exploitation

Enfin, on peut définir un paramètre pour tenter une mesure d'un système de gestion de base de données. on associera à une structure le rapport entre le volume en caractères des données brutes et le volume de la

base de données. Ce rapport est toujours inférieur à 1. Un autre rapport est également significatif de la structure des fichiers, c'est le taux de remplissage, rapport du volume de mémoire occupée effectivement au volume de mémoire réservée.

Chapitre II

BANQUES DE DONNEES

2.1. Notion de banque de données.

On appellera banque de données un ensemble composé par :

- une base de données,
- un système de gestion de la base de données,
- un ensemble de programmes d'exploitation de la base de données.

Cet ensemble de programmes est constitué d'éléments divers ayant des destinations bien définies. On distingue :

- les programmes d'interface, permettant les échanges d'informations des fichiers entre le système donné et les autres systèmes,
- un sous-système d'exploitation destiné à une application restreinte et définie,
- un groupe de moyens divers permettant la mise en oeuvre de l'information stockée par une classe bien définie de programmesutilisateurs,
- un groupe de moyens d'évaluations.

- Les programmes d'interface:

Peu de banques de données présentent, à l'heure actuelle, dans leur système, de tels programmes. Cependant, la multiplication des systèmes obligera les créateurs à se soucier du problème, du seul point de vue, d'ailleurs, du coût. L'exploitation d'une banque de données est couteuse au niveau de la saisie de l'information, de la mise en forme, de la vérification. Ces coûts ne peuvent être abaissés que si les opérations ne sont faites qu'une fois. Dès lors, il convient, pour un certain nombre de banques de données, de procéder à des échanges d'informations.

Deux cas encore sont à distinguer : échange "on line", échange de fichiers sur support informatique.

L'échange "on line" d'informations s'envisage dans le cas où les banques de données sont implantées sur des ordinateurs inter-connectés.

L'échange de fichiers entre système est un problème difficile à résoudre dans la pratique autrement que par échanges de fichiers cartes ou de fichiers construits sous le même operating system.

Les premières difficultés se situent au niveau hardware et mode d'enregistrements. Ensuite, la structure logique d'un fichier étant liée à une base de données, elle n'est pas souvent compatible avec un autre système. Aussi, est-il souhaitable, en un premier temps, de définir des structures et des formats d'échanges. Le problème a été abordé dans le domaine de la documentation automatique par la création du format MARC 2. Un autre effort a été tenté au Chemical Abstract qui distribue ses fichiers dans un format standard (SDF) avec un certain nombre de sous-programmes permettant d'accéder à l'information.

- Le sous-système d'exploitation:

Le sous-système d'exploitation peut se présenter sous deux formes :

1 - il s'agit d'un module d'un système général de banque de données, indépendant de l'application envisagée.

Les objectifs visés seront :

- accès à l'information en temps réel,
- possibilité de calculs et d'édition, portant sur les informations extraites.

L'accès à l'information sera rendu possible par l'utilisation d'un langage qui permet la sélection des articles , ou leur comptage. Dans le cas d'un langage conversationnel, il y a lieu de considérer trois phases dans la mise en oeuvre de ce langage:

- Elaboration de la question Il s'agit de permettre à l'utilisateur d'acquérir l'information représentée par la description des fichiers et des articles

- Mise au point de la question du point de vue syntaxique et sémantique.

- Définition de la sortie.

Le langage est souvent étendu pour permettre des calculs sur les éléments recherchés et la sortie des résultats. En particulier, il permet la définition de mise en page, de formats. Enfin, le langage inclut des possibilités de réorganisation des données extraites.

Ce langage est aussi souvent utilisé pour la mise à jour "on line" qui n'apparaît que comme une des fonctions de celui-ci.

2 - La deuxième forme du sous-système d'exploitation est celle prise par un sous-système adapté à une application déterminée. La nature des informations traitées dans l'application prend toute son importance dans la définition des objectifs aisés par le sous-système et des moyens mis en oeuvre pour y parvenir. Cependant, les objectifs restent :

- accès à l'information en temps réel le plus souvent,
- possibilité de traitements et d'éditions.

Les possibilités de traitements sont souvent plus importantes que dans le cas précédent, tout en étant moins grandes. Pour parvenir à ces objectifs, l'utilisateur a, à sa disposition, dans la plupart des cas, un langage d'accès aux données qui comporte un certain nombre de fonctions orientées dans le sens de l'application.

- Les programmes d'applications.

Ces programmes se déroulent en dehors de la banque de données, mais en utilisant des modules qui peuvent appartenir au sous-système d'exploitation. Ces modules permettent :

- l'accès aux informations,
- la définition et la restructuration de sous-fichiers.

Ces programmes seront souvent utilisés sous le mode "batch processing" et peuvent constituer une bibliothèque de programmes d'applications relevant de domaines divers.

- Les évaluations.

Il s'agit pour le système de connaître son état à chaque instant et de pouvoir s'adapter.

Il faut noter la présence de programmes permettant d'établir une statistique des mouvements et des types de questions posées, ou encore un programme permettant d'adapter la valeur de certains paramètres généraux du système à la nature de l'utilisateur.

Ces programmes sont présents dans les systèmes généraux ou dans

les systèmes qui assurent un partage des fichiers entre utilisateurs, très divers quant à la nature des terminaux employés, des tâches soumises, etc...

2.2. - Exemples de système de banque de données

2.2.1. - Le système MISFIIT.

Le système MISFIIT, conçu par le CETE d'Aix en Provence, est un système d'interrogation et de gestion de base de données dont les caractéristiques sont les suivantes :

- MISFIIT fonctionne en temps réel.
- L'interrogation et la mise à jour se font à l'aide d'un langage conversationnel qui n'est pas un langage de programmation classique.
- L'interrogation peut être menée simultanément par plusieurs utilisateurs.
- La recherche se fait sur n'importe quelle partie de l'article.
- L'utilisateur dispose de statistiques de pré-recherches.

Un fichier de données, à articles de longueur variable, est accessible par consultation d'un fichier inversé à indexation totale (F I I T). Tout élément, groupe qui peut avoir plusieurs occurrences dans le même article, constitue une entrée du fichier.

Les caractéristiques des entrées composent le premier élément du fichier F I I T, le lexique. Celui-ci est composé d'une table des entrées associant un nom d'entrée à un pointeur vers la table structure donnant pour chaque entrée l'appartenance et le type de l'entrée. Trois autres tables indicées par tous les groupes possibles d'un article donnent :

- le nombre d'entrées pour le groupe,
- le niveau dans l'arborescence des groupes,
- le "père" de ce groupe.

Une quatrième table fournit :

- le format de la valeur d'une entrée,
- la plage des valeurs possibles,
- l'unité, etc...

Le deuxième élément constituant du fichier FIIT est formé par un ensemble de listes appelées listes valeurs. A chaque entrée, est associée une liste des valeurs prises par celle-ci. La liste se décompose en un "directory" permettant de localiser une valeur ou un ensemble de pages de valeurs. A chaque valeur est associé, selon le cas, un pointeur vers une liste "ordonnée" ou le "grand mot" associé. Le "grand mot" est un ensemble de N bits, où le bit J est positionné à 1 si l'enregistrement J a pris la valeur associée.

Les listes ordonnées font correspondre à un couple entrée-valeur la liste des enregistrements où la valeur est prise. Les pointeurs pointent vers les listes "enregistrement" qui donnent les adresses physiques des enregistrements et leur longueur. Une liste ordonnée est utilisée à la place du grand mot dans le cas où elle occuppe moins de place que celui-ci.

Les parties "filiations" et "données" du fichier FIIT permettent d'associer à chaque enregistrement la hiérarchie qui le décrit.

Un élément de "filiations" se rapporte à un groupe et est constitué par :

- un chaînage vers le groupe de niveau inférieur,
- un chaînage vers une nouvelle occurrence du groupe,
- un numéro,
- un pointeur dans la partie "données".

La partie "données" donne pour chaque groupe l'ensemble des valeurs prises par les éléments du groupe. Chaque valeur d'une entrée est repérée dans cette partie par le numéro du groupe et sa position dans celui-ci.

La création du fichier FIIT, sur mémoire à accès direct, est assurée par un programme "batch".

La mise à jour peut s'opérer à partir d'un terminal ou à partir de bordereaux perforés. Une mise à jour à partir du terminal peut être faite en temps réel ou temps différé.

En raison de la complexité des relations entre tables,

une modification de structure d'un article est pratiquement impossible et nécessite une nouvelle création.

La mise à jour s'appuie sur le langage d'interrogation qui permet de sélectionner les articles atteints par l'opération en cours. La modification d'une valeur dans un article entraîne suivant les cas des suppressions ou ajouts dans diverses parties de FIIT ou encore la création de chaînages. La suppression d'un article se ramène à une suppression dans la liste "enregistrement".

La protection de l'information se situe au niveau du fichier, au niveau des entrées.

L'interrogation est assurée au moyen d'un langage conversationnel qui permet :

- d'élaborer la question,
- d'effectuer les corrections syntaxiques et sémantiques nécessaires,
- de définir la sortie.

Les fonctions assurées dans un premier temps sont : sélection, comptage, mise à jour.

Les commandes "DECRIRE" et "MONTRER" permettent à l'utilisateur d'accéder aux entrées (descriptions de la structure de l'article), et aux valeurs associées.

Trois instructions sont groupées :

- SELECTIONNER,
- COMPTER,
- METTRE A JOUR.

Ces instructions n'entraînent pas l'accès aux articles eux-mêmes, mais elles provoquent la consultation du fichier FIIT pour connaître les numéros d'enregistrements.

D'autres instructions permettent alors l'extraction des articles et leur affichage ou sortie sur imprimante ou bande magnétique en format standard.

L'instruction fondamentale de recherche est "SELECTIONNER" E (ei) 'SI' E (ej). Les ej servent à définir les critères de selections. Les critèressimples sont soit arithmétiques et construits avec les opérateurs classiques, soit le critère simple de possession.

Les critères simples peuvent être utilisés pour définir un critère, par opérateurs booléens (ET, OU, ET NON, OU NON).

Les critères multiples sont des sommes booléennes de critères simples.

Les ei sont des entrées qui définissent le niveau de sélection dans l'article et qui doivent être sélectionnées.

L'interprétation du langage se fait en 3 phases :

1 - analyse de la question. Cette analyse se fait par codage de la question et constitution d'une chaîne post-fixée.
Un certain nombre d'erreurs sémantiques syntaxiques sont détectées.

2 - Recherche des enregistrements à consulter. Elle s'opère par interprétation de la chaîne. Elle provoque le marquage des enregistrements à consulter.

3 - Recherche des informations dans les enregistrements repérés.

MISFIIT est un système implanté sur l'ordinateur IBM 360/50 et sera complété dans un avenir proche.

La structure du fichier FIIT est complexe et orientée vers l'amélioration des performances de recherches pour le programme d'interrogation uniquement. Celui-ci peut être complété par la possibilité d'enchaîner la réponse à un programme de traitement de l'information sélectionnée.

2.2.2. - Le Projet SOCRATE.

Le projet SOCRATE développé à l'IMAG est un ensemble de spécifications et de techniques pour un système de banque de données.

Le système fonctionne sous trois modes principaux :

- le mode "définition de données". Un langage de définition permet à l'utilisateur de définir des structures de données complexes.

- le mode "manipulation de données". L'utilisateur peut créer, supprimer, mettre à jour, des données dans la structure définie.

- Le mode "récupération". Un langage d'interrogation permet d'exprimer des questions concernant les données. Ce langage inclut, outre des possibilités de sélections, des moyens de traitements par l'intermédiaire d'un langage algorithmique simple

Les trois modes peuvent être activés simultanément par un grand nombre d'utilisateurs travaillant ou non sur les mêmes fichiers

A ces trois modes, s'ajoutent des modes de fonctionnement spéciaux :

- le mode "Statistique" : la réponse à une interrogation n'est pas constituée par les éléments du fichier, mais est le résultat de ventilations et de totalisations.

- le mode "mise à jour spontanée" : il s'agit d'un fonctionnement permettant aux fichiers d'évoluer indépendamment des utilisateurs en fonction du temps et des actions extérieures ; le système détecte de lui-même les relations qui sont à modifier ; celles-ci peuvent être dues à l'évolution de la structure des fichiers ou à la nature sémantique des informations.

- Le mode "Modification de Structure" : ce mode permet la redéfinition de la structure d'un fichier du système.

- Le mode "Documentation" : ce mode permet à l'utilisateur d'obtenir toutes les informations utiles pour poser une question.

Le système est conçu pour en permettre l'utilisation "on line" Il met en oeuvre des moyens de protection en prévision d'un mauvais fonctionnement du système. Il peut limiter l'accès à tout ou partie des fichiers.

Il existe un interface permettant à l'utilisateur d'inclure parmi les fichiers du système, des fichiers externes de type "Cobol".

Le système requiert l'utilisation de 3 langages :

- le langage de définition de structure,
- le langage de citation (interrogation),
- le langage de commandes.

Le langage de définition permet la rédaction d'un programme qui construit la structure et associe à chaque élément en groupe, un nom. Les informations sont réparties en 4 types :

- chaînes de caractères alpha-numériques,
- liste de valeurs,
- valeur numérique appartenant à un intervalle,
- texte.

A chaque élément sont associés un identificateur et un type.

Un groupe est décrit comme un bloc qui est associé à un identificateur. Un groupe peut être répétitif et se réaliser à l'intérieur d'un groupe identique. Il constitue une entité. Le langage introduit une structure syntaxique qui permet la description de structures arborescentes. Il permet aussi les descriptions d'entités conditionnelles définies par référence à la valeur d'un élément. Il est possible de définir d'autres références que celle d'appartenance.

Une structure de données sous forme de graphe est complètement décrite par le langage. Le langage de citation permet de désigner un ensemble de réalisation d'un certain objet. Ceci peut être fait, au point de vue formel, par rapport à la définition de la structure. La citation s'exprime par des "conditions" introduites par des opérateurs. La citation est complètement

hiérarchisée pour définir une entité, un groupe ou un élément en définissant tous les niveaux.

Les conditions exprimées par les opérateurs peuvent être combinées au moyen d'opérateurs booléens (ET, OU).

Deux conditions "existentielles" sont introduites : EXISTE et AUCUN sont des opérateurs binaires placés en tête d'une citation dans une condition. Deux quantificateurs UN et TOUT ont été ajoutés pour alléger l'expression.

Le langage de commande permet à l'utilisateur d'activer des fonctions du système telles que (A) addition d'entité, (C) création, (E) effacement, (M) mise à jour, (I) interrogation, (N) numération, (S) surveillance, (P) groupement d'actions en boucles.

Cette partie du langage est conversationnelle. Le langage permet des actions sur un ensemble de fichiers.

L'utilisateur peut alléger le langage en utilisant un macro-générateur qui, modifiant le langage externe, génère un langage interne identique.

Le système considère deux espaces : l'espace virtuel et l'espace réel.

Le fichier est implanté dans l'espace virtuel vis-à-vis de toutes les opérations Un module permet d'associer à une adresse virtuelle une adresse physique.

Lespace virtuel est segmenté. Un segment contient de l'information ; par opposition, un trou n'en contient pas. Chaque segment est divisé en quatre parties :

- une partie "service" réalisant un chaînage et donnant diverses tailles.
- une partie "dictionnaire" à pointeurs vers les segments "fils",
- une partie "valeur" qui contient l'information.
- Une partie "libre".

Les "trous" sont chaînés entre eux.

L'espace réel est supporté en partie par la mémoire centrale (plan primaire), en partie par une mémoire à accès direct (plan secondaire). Il est découpé en pages. Les deux plans échangent des pages.

La relation adresse virtuelle-adresse réelle peut se faire par calcul d'une adresse dans le plan primaire : par hash -code de l'adresse virtuelle. L'allocation de mémoire virtuelle à une information se fait par consultation d'un dictionnaire dont l'entrée est l'adresse et qui marque les places occupées. En cas de collision, on utilise une fonction de déplacement de l'entrée dans le dictionnaire.

D'autres méthodes d'établissement de relations entre espace virtuel et espace réel peuvent être mises en jeu.

Dans les fichiers, l'information est condensée au maximum par des méthodes de compactage.

L'adresse virtuelle d'un objet dont est donnée la réalisation, est calculée par diverses méthodes dépendantes de l'organisation adoptée.

La multiprogrammation est réalisée par un "scheduler" qui a pour rôle d'activer 3 "processus" indépendants :

- le processus "ligne" assure la transmission des informations entre le système et les consoles.

- le processus "page" est chargé des transmissions de pages entre la mémoire secondaire et la mémoire principale .

- Le processus "système" réalise toutes les fonctions de la banque.

Les caractéristiques les plus importantes du système SOCRATE sont, au travers des diverses méthodes et techniques proposées, les suivantes :

- fonctionnement général "on line",
- un seul langage assure :
  - . la définition des données dans des structures générales,
  - . la création,
  - . la mise à jour,
  - . l'interrogation,
  - . le traitement des applications.
- L'accès peut être multiple.
- Les fichiers, sur mémoire à accès direct, sont gérés dans un espace réel en référence à un espace virtuel.
- L'adresse d'une information est calculée dans l'espace virtuel
- Les paramètres principaux du système sont ajustables. Ils peuvent évoluer sous l'action de programmes.
- La protection de l'information est assurée.
- Le système peut prendre en compte un fichier créé par un autre système (Cobol).

Les deux systèmes, succinctement décrits précédemment, entrent dans la classe des systèmes généraux de banques de données indépendants des applications. Ils sont utilisables pour la construction et l'exploitation de banques de données dans les domaines les plus divers.

2.3. - Caractérisation d'une banque de données.

La caractérisation d'une banque de données est double : elle concerne la base de données et l'exploitation de celle-ci.

L'interrogation et l'exploitation de la banque de données sont caractérisées par leur mode de fonctionnement : "batch", ou "on line". Dans ce dernier cas, à une utilisation par un seul utilisateur pour l'interrogation, on oppose l'utilisation partagée entre plusieurs utilisateurs, tant pour la consultation que pour la mise à jour "on line". La nature du superviseur d'entrée-sortie est également un critère de caractérisation Ce sous-système assure, en plus de l'interrogation, diverses fonctions que l'utilisateur peut trouver :

- possibilités de calculs et traitements divers,
- génération de rapport sur différents périphériques,
- mise en forme des sorties,
- protection de l'information.

On peut aussi caractériser l'interrogation par le langage mis en oeuvre, en s'intéressant aux points suivants :

- langage de programmation ou langage spécifique,
- performances,
- états conversationnels,
- extensibilité,
- fonction documentaire : il s'agit de donner la possibilité à l'utilisateur d'accéder à la description et à la structure des fichiers.

L'exploitation de la base de données par les programmes utilisateurs est caractérisée :

- par la nature des modules de recherches et d'accès utilisés (sous-programmes, macro-instructions, éléments de langages, etc ...),

- par le langage d'écriture de ces programmes d'applications,

- par la possibilité d'accès simultanés aux fichiers.

Les programmes d'interface sont caractérisés par :

- leur degré de généralité,

- leur mode d'exploitation,

- la nature des fichiers mis en oeuvre,

- le langage de programmation utilisé.

La banque de données sera aussi caractérisée par les paramètres généraux du système : taille de mémoires, nombre de terminaux, nature de ceux-ci, etc ... et les possibilités d'adapter ces paramètres à tous les cas d'exploitations.

Enfin, une dernière caractérisation difficile à obtenir est la charge que représente pour l'ordinateur l'exploitation de la banque de données.

Chapitre III

SCIENCES HUMAINES ET BANQUES DE DONNEES

3.1. - Les informations archivées.

On peut distinguer, parmi les informations archivées pour l'étude et la recherche dans les sciences humaines, deux genres d'informations suivant la nature de l'entité représentant l' "objet".

L'entité est descriptible par des éléments quantitatifs ponctuels. L'ensemble forme un fichier "ponctuel". Ainsi, un fichier d'établissements industriels et commerciaux est un fichier ponctuel. L'entité "établissement" est parfaitement définie et elle est caractérisée et décrite par des éléments qui prennent des valeurs numériques ou alphanumériques, le plus souvent dans un ensemble fini . On peut définir, suivant les cas, sur de telles entités, des opérations de comptage, ventilations et totalisations.

A de tels fichiers, s'opposent les fichiers agrégés dont les entités sont formées d'éléments de nature tabulaire et d'éléments divers dont l'un sert de clé d'accès, ainsi que les fichiers de tableaux des recensements, les tables de contingences d'une enquête. On peut définir de multiples opérations entre de telles entités et également sur l'ensemble des éléments de ces entités.

La construction d'une entité pour la constitution des fichiers est dominée par la définition des éléments qui définissent cette entité et dont les valeurs la caractérisent. Cette définition est décrite dans un langage. Dans certains cas de banque de données, l'utilisateur aura accès à l'information en utilisant un langage d'interrogations, dont le vocabulaire lui est imposé par le système. Ces éléments du vocabulaire lui permettent de désigner les éléments des entités qu'il cherche à repérer ou atteindre.

L'importance du choix du vocabulaire est alors évidente. De la "régularité" de la définition des éléments et des désignations, dépend la satisfaction de l'utilisateur. En effet, l'usage d'un vocabulaire inadéquat peut égarer l'utilisateur dans ses recherches d'informations.

Pour permettre l'utilisation de tels systèmes à une large classe d'utilisateurs, on utilise souvent deux langages. Le premier sert à l'interrogation et peut être très proche du langage courant. Le second est un langage de description des entités. Un dictionnaire permet alors de désigner des éléments par le langage d'interrogation. Ce dictionnaire assure le passage entre les deux langages. Il le fait souvent par l'intermédiaire d'un thésaurus. Un thésaurus est un ensemble de relations entre concepts. Il permet de structurer ceux-ci, souvent en arbre, quelquefois en graphe. L'utilisation d'un thésaurus est faite pour assurer une meilleure traduction d'une interrogation, au niveau sémantique. L'interrogation générée dans le langage de description permet alors une désignation par concept associé en vocabulaire du langage de description. La construction d'un thésaurus ne relève pas de l'informaticien, mais des spécialistes en sciences humaines. Le problème est alors connexe de celui de l'indexation des documents qui peut exiger pour l'exploitation du système documentaire qui en résulte, l'utilisation d'un thésaurus.

Le manque de "régularité" des définitions des objets constitués par des éléments d'informations, dont les désignations peuvent sémantiquement évoluer d'un contexe à l'autre, entraîne la construction de fichiers de données annexes. L'importance de ces fichiers peut être comparable à celle des fichiers de données proprement dites. Leur organisation peut être très complexe et donc être un facteur d'allongement de temps de réponse dans une banque de données.

La définition des données est influencée par deux facteurs :

- diversité des utilisateurs,

- diversité des sources.

Les utilisateurs sont différents les uns des autres parce que l'usage d'une banque de données, relevant par exemple des sciences économiques, ne peut être réservé à une certaine classe d'économistes, parce que, également certains phénomènes sont approchés et donc les informations qui les caractérisent sont désignées par des méthodes conceptuelles différentes et des langages qui relèvent de celles-ci. Ainsi, on voit apparaître des banques de données économiques spécialisées suivant l'utilisation qui en sera faite.

La diversité des sources d'information ajoute à la confusion. Telle information économique désignée par tel organisme collecteur d'informations est toute différente de celle désignée de façon identique par l'organisme pour qui elle résulte d'un acte de gestion. Le problème est également posé en documentation automatique.

3.2. - Problèmes de l'archivage.

L'un des problèmes de l'archivage est un problème de conservation. L'enregistrement magnétique des informations est, qu'il s'agisse de textes, d'informations statistiques, etc ..., une des solutions qui peut remplacer l'archivage sous forme de documents imprimés Ceux-ci sont trop volumineux, difficiles à reproduire, à manipuler, à conserver. La conservation de fichiers sous formes lisibles par l'ordinateur est un problème qui commence à se poser, par exemple en Amérique, dans les DATA ARCHIVE CENTERS. Le premier point est la conservation physique de l'information sur son support. Ceci peut être la cause de grands frais. Le second est celle de la conservation structurelle et sémantique. Il s'agit de pouvoir décoder toute cette information au niveau software, puis au niveau utilisation. A cet égard, l'enregistrement des structures et des formats est déjà une première mesure. Le problème est encore trop neuf pour que des solutions satisfaisantes puissent se dégager.

3.3. - Les hiérarchies de l'information.

L'utilisateur d'une banque de données relevant des sciences humaines s'adresse, tantôt à la fonction Archivage, tantôt à la fonction Traitement des Données. Il semble que l'on puisse mettre ces deux fonctions sur deux plans séparés du point de vue de l'utilisation.

L'utilisation des archives suppose d'abord une recherche à l'intérieur de celles-ci. Du fait de l'extrême abondance de l'information, cette recherche, souvent menée "à l'aveuglette", peut conduire à des résultats volumineux. Il s'ensuit que le coût d'exploitation d'un système d'interrogation "on line" permettant d'accéder à l'information archivée elle-même devient prohibitif. Cependant, il est nécessaire de pouvoir assurer une fonction documentaire à propos des archives et ce, en temps réel, en mode conversationnel.

La seconde fonction, utilisée assez souvent, est le traitement des données de façon plus ou moins poussée, et, en tout cas, adaptée aux fichiers abordés. Dans cette hypothèse, il semble donc nécessaire que l'utilisateur puisse le faire également en temps réel et, si possible, de façon simple et conversationnelle. Auparavant, le système aura dû extraire des archives une information sélectionnée par l'utilisateur.

On pourrait donc imaginer une banque de données économiques par exemple fonctionnant de la manière suivante : la base de données est à deux niveaux : le premier représente les archives proprement dites ; elles ne sont pas accessibles directement par l'utilisateur. Le deuxième niveau est constitué par les données extraites des archives sur demande, après sélection, sur lesquelles des traitement peuvent être effectués. Le deuxième niveau contient également les informations qui permettent à la banque de données de satisfaire la fonction documentaire couvrant les archives. A la gestion de la base de données, s'ajoute alors l'action de passage des informations en archive vers le deuxième niveau. Cette action, engagée par l'utilisateur selon

les choix, peut être faite en temps différé. En effet, on peut estimer que cette action effectuée dans des délais voisins de la journée, n'allonge pas trop les temps de réponses globaux du système dans la plupart des cas. Un temps de réponse très bref ne sera exigé que de systèmes où les informations doivent être prises en compte pour la régulation d'un processus dont le déroulement se fait sur un rythme relativement rapide. Ce peut être le cas de certaines gestions, de systèmes de réservations de places, d'applications médicales, etc... Ce ne sera pas, bien souvent, le cas des applications de sciences humaines Cependant, il peut être utile de développer, au deuxième niveau, une stratégie de l'information pour que la plupart des besoins rapides d'informations traitables puissent être satisfaits. La base de données doit pouvoir alors évoluer suivant l'usage qui en sera fait. En effet, il existe notamment dans des banques de données économiques, des informations qui sont très souvent utilisées par rapport à d'autres informations peu utilisées. La stratégie consiste alors à maintenir cette information utilisée "on line"

Le choix des domaines couverts par la banque de données est fonction du volume d'informations à archiver et des ressources. C'est pourquoi, il peut arriver que l'on limite le champ de la banque Cependant, les limites sont bien souvent floues. Il semble alors préférable, au lieu de constituer des archives sur des supports coûteux, de stocker les informations qui permettent de localiser à l'extérieur de la base de données, les données qui, n'entrant pas dans le champ couvert, n'en sont pas moins très proches. Ainsi, se constituera un troisième niveau qui est formé par l'information sur les données externes que des interfaces pourront intégrer aux archives.

Pour l'utilisateur, la banque de données permet alors de rechercher des informations dans un champ très étendu, de disposer d'archives importantes, et de traiter de façon performante des données extraites des archives.

3.4. - Un problème classique : le dépouillement statistiques.

L'un des cas les plus fréquents d'utilisation de l'informatique est le dépouillement d'enquêtes. Il s'agit d'un dépouillement de fichier complété par des traitements statistiques qui peuvent s'étendre d'un simple calcul de moyenne à une analyse de variance en autres "modèles" plus complexes. Ce problème, bien qu'indépendant des banques de données, peut toutefois dans le cas des sciences humaines y être rattaché. Il peut être en effet résolu par l'utilisation d'un programme externe à la banque de données mais le fichier d'enquête appartient à la base de données. Cependant, dans le cas de banques de données particulières, un traitement spécial de dépouillement est prévu.

Il existe de nombreux packages permettant en "batch" à l'utilisateur de construire par ventilation des tables de contingences à plusieurs dimensions et d'éditer des tableaux. L'utilisation d'un langage, généralement interprété, est requise pour :

- la définition du fichier (phase de déclaration),
- la construction des critères de sélections,
- la définition des tables de contingences,
- les éditions de tableaux.

Le langage est plus ou moins élaboré, suivant les cas. Le fichier est souvent de nature séquentielle.

Des travaux, menés pour les sociologues de la Faculté des Lettres à Lille et les économistes de la DATAR, ont permis la création de deux programmes d'inspiration similaire. La méthode consiste d'abord à créer, à partir du fichier d'enquête vérifié par un programme spécifique, un fichier

pour exploitation de structure simple et linéaire. Ce fichier est recréé chaque fois que l'on définit de nouvelles variables de croisement en ajoutant des instructions nouvelles, instructions faciles à insérer puisqu'elles sont uniquement arithmétiques et que les programmes sont en FØRTRAN ou en ALGOL. Le fichier obtenu est alors exploité. La définition des tableaux des éléments de mises en page (libellés, formats, principalement) est introduite par cartes dans un cadre très rigide.

Cette méthode est évidemment limitée dans ses applications mais parait convenir dans la plupart des cas courants lorsqu'on dispose d'un ordinateur à bandes.

On peut reposer le problème des dépouillements d'enquêtes dans le contexte des banques de données. L'utilisateur devrait pouvoir disposer, pour traiter des fichiers qui ne soient pas trop volumineux en temps réel, d'un langage qui puisse assurer :

- la création,
- la construction de tableaux,
- le traitement statistique des tableaux,
- l'édition des résultats "off-line",
- la mise en oeuvre de méthodes d'analyses statistiques,
- la sortie des résultats de celles-ci, sur supports divers (télétype, display cathodique),
- l'établissement de relations entre fichiers divers.

Il faut également que le langage mis au point soit simple pour une utilisation par des économistes, sociologues, psychologues, que les interfaces avec les langages de programmation soient développés afin de permettre l'exploitation des fichiers en résultats intermédiaires par des programmes plus spécialisés.

3.5. - Conclusion.

Les problèmes soulevés par la création et l'exploitation de banque de données pour lès sciences humaines, relèvent à la fois de l'informatique et des disciplines de ces sciences.

Les banques de données de type système de documentation automatique, sont les plus nombreuses. Elles peuvent être utilisées dans bien des cas, tous ceux où l'utilisateur n'interroge la banque que pour connaître une référence à un objet, par exemple livres, articles, tableaux statistiques considérés comme documents, cartes de géographie, objets archéologiques. D'autres, plus rares, mettront en jeu divers fichiers (enquêtes, recensements sous formes ponctuelles ou tabulées, résultats électoraux etc ...). L'utilisateur pourra alors traiter les informations après l'interrogation.

Partie II

SYSTEME D'ACCUMULATION DES DONNEES

Le système d'accumulation des données fut créé par le Professeur J.P. TRYSTRAM pour répondre aux besoins de la Délégation à l'Aménagement du Territoire et à l'action régionale en matière de données statistiques localisées. Le premier travail qui fut fait consistait en un "repérage" : il s'agissait de dresser un inventaire des ressources en la matière. Pour cela, une petite équipe d'économistes eut à dépouiller des documents où se trouvait depuis 1955 l'information statistique. Le résultat fut surprenant : il est possible de trouver plus de 200 000 données pour chacun des départements français et donc environ autant pour les régions de programmes. L'importance de cet ensemble de données ne doit pas faire illusion. Peu d'entre elles sont directement utilisables. Cependant, la Délégation à l'Aménagement du Territoire et à l'action régionale entreprenait de publier un document appelé "Statistiques et Indicateurs des Régions Françaises" qui avait pour objectif de présenter environ 700 données sur les 21 régions de programmes avec pour certaines leur évolution depuis dix ans au maximum. Ces données provenaient d'un traitement sur un ensemble d'environ 2500 données élémentaires. Aussi le recours à l'informatique s'imposa-t'-il, d'autant plus qu'on pouvait alors envisager la création d'une banque de données répondant aux besoins d'informations statistiques localisées nécessaires à l'action courante de la DATAR.

On appellera série un ensemble de données localisées, c'est à dire liées à un ensemble d'unités de territoire issues d'un découpage de celui-ci, observées ou recueillies pour une ou plusieurs années. On appelera série-année les valeurs d'une série pour une année donnée. Le SAD constitua donc d'une part des fichiers où se trouvaient enregistrées des séries sous forme de série-année, d'autre part un fichier documentaire qui résultait du repérage de l'information. Le SAD avait alors pour objectif de résoudre deux problèmes. Le premier est de pouvoir "retrouver" dans ses fichiers les séries permettant à un utilisateur de traiter son problème. Il s'agit ici d'un problème de type documentaire. L'utilisateur ne connait pas à priori la nature des données enregistrées et désire connaître les données relatives à un phénomène ou une situation déterminés.

Le second problème est de pouvoir traiter les données "retrouvées", soit que, dans l'état où elles se trouvent, elles se révèlent inutilisables directement, soit que l'utilisateur désire effectuer un traitement particulier.

Le système SAD I qui va être décrit résolvait partiellement le premier problème ; le second fut complètement résolu. Cependant les besoins croissaient, le coût de stockage de l'information également ; la lourdeur d'un système "batch" était telle que, à partir de cette expérience, on décidait de créer un nouveau système élargissant le premier objectif à toute l'information disponible ; le second objectif devait être atteint grâce au traitement à distance et sur le mode conversationnel. Le système SAD II devra atteindre ces objectifs et devenir ainsi une banque de données statistiques localisées accessibles sans intermédiaire à tout utilisateur, même éloigné du centre de calcul.

Un tel système ne se fabrique pas rapidement et il s'imposait de faire quelques expériences dans le domaine de l'utilisation. Aussi le SAD mit-il au point une version restreinte de SAD II, fonctionnant sous un système de Time-Sharing commercial. Cette version, complètement opérationnelle, sera décrite surtout pour les conclusions qui ont pu en être tirées.

Chapitre I

LE SYSTEME SAD I

1 - Problèmes Généraux

L'information traitée par le système est constituée principalement par :

1) Des séries statistiques départementales et régionales

2) Les descriptions de ces séries.

Le traitement des séries a donné naissance à un système qui a fonctionné de 1967 à 1971 sur CII 90-80. Ce système réalisé en plusieurs étapes a été complété et exploité par nos soins. Aussi sera-t-il décrit succinctement pour montrer les avantages et inconvénients qu'il a présentés et comment il est devenu nécessaire d'en envisager une refonte complète.

Le traitement des descriptions de séries, appelées "titres", a été réalisé comme expérience sur l'ordinateur BGE M40 du laboratoire de calcul de la Faculté des Sciences de Lille. Le programme ne fut jamais exploité pour des raisons évidentes d'éloignement. Cependant, il permit de tester la validité "opérationnelle" des concepts qui présidèrent à son élaboration.

1.1 - Le Traitement Statistique

Le problème majeur soulevé par l'exploitation reste la gestion des fichiers de données. Le premier objectif à atteindre est la facilité d'enregistrer des données sur bandes magnétiques. Le second est de pouvoir faire des mises à jour massives, fiables et répétées sans obérer les délais. Enfin, l'un des buts de l'opération était de pouvoir construire à partir des fichiers de séries statistiques, des tableaux statistiques de présentation relativement normalisée dont le contenu devait être obtenu, au moyen d'un langage simple, par certains traitements des données élémentaires. Ces objectifs trouvent leur justification dans le fait qu'il fallait faire fonctionner le système en confiant les tâches d'exploitation et de gestion à des spécialistes non informaticiens de l'information économique, avec un minimum de délais pour l'utilisateur dont les demandes d'information pouvaient être relativement importantes.

1.2 - Recherche de titres

Devant le mombre très important de données statistiques existantes, il a été nécessaire de faire un choix pour la constitution des fichiers. D'autre part, l'utilisateur de données statistiques ignore bien souvent la richesse d'informations et leur degré de finesse dans l'observation. Il fallait donc pouvoir lui offrir un instrument de recherche sélective de titres de séries utiles, parmi l'ensemble de tous les titres. Cette recherche est fondée sur l'utilisation d'un langage de description des données statistiques. Le langage permet de structurer le fichier des titres et de coder ceux-ci de manière économique et utile pour la recherche.

Le programme mis au point constitue l'exploitation des fichiers de titres. L'expérience, bien qu'intéressante et concluante, fut cependant arrêtée en partie à cause du fait qu'elle se déroulait éloignée du lieu d'implantation du premier système. Aussi, les questions de mise-à-jour ne furent-elles pas abordées. Cependant, elle permit de prolonger la recherche dans la direction prise et de mettre au point définitivement le langage de description des données statistiques et d'envisager, ce qui était impossible à Lille, un système intégré de recherche et traitement de séries sur le mode conversationnel.

Les deux parties du système SAD I ont donc été réalisées sur des ordinateurs de la deuxième génération exploités en monoprogrammation. Les langages de programmation employés, FORTRAN IV et ALGOL ne rendaient pas faciles leur intégration au sein d'un système unique. Les limites d'emploi d'une part et les besoins toujours croissants d'autre part ont imposé l'élargissement des objectifs. Cependant, on a bien vu au cours de cette expérience qu'un nouveau système devait s'appuyer sur des techniques conversationnelles permettant au sous-système documentaire, dont la nature est spécifique du problème posé, de travailler efficacement et permettant d'exploiter et de gérer, de façon économique et rapide, les fichiers de données statistiques.

2 - Organisation de SAD I (Traitement numérique)

Le système repose sur la gestion et l'exploitation de deux fichiers.

2.1. - Le fichier départemental (FD)

Ce fichier sur bande magnétique contient l'information au niveau départemental. Il est constitué par des série-année départementales. Une série - année départementale est donc une liste de 90 ou 91 (95 ou 96) nombres selon qu'il s'agit d'une série sommable (c'est-à-dire que la somme des valeurs de la série est la valeur correspondante de la série pour le niveau "France Entière"), ou non pour un découpage du territoire français en 90 ou 95 départements.

Une série-année correspond à un article du fichier. Cette série-année est enregistrée sous forme d'entiers grâce à l'introduction du facteur d'échelle. Ce nombre, figurant dans l'article, est l'exposant de la puissance de 10 par lequel il faut multiplier une valeur de la série-année pour obtenir la valeur réelle enregistrée. Déterminé à l'enregistrement par programme, il permet également de tronquer la partie significative à droite afin de pouvoir stocker la valeur sous la forme entière la plus proche de 2^{23-1} (valeur du mot à 24 bits du CII 90.80). Ce facteur d'échelle est codé avant enregistrement de la façon suivante : soit x la valeur à enregistrer du facteur d'échelle pour une série donnée, et soit y le code correspondant, on passe de x à y par la transformation :

Si $x < 0$ alors $y := 100 + x$ sinon $y := x$;
Si série sommable alors $y := y + 100$;

Le code y est alors enregistré. Ce code permet donc de savoir si la série-année que représente l'article est sommable ou non et de calculer le facteur d'échelle. On remarquera que celui-ci est supposé inférieur, en valeur absolue à 50 de telle sorte que connaissant y on tire x par :

Si y > 100 alors y : = y - 100 ;
Si y > 50 alors x : = y - 100 sinon x : = y ;

Ce code de facteur d'échelle sera désigné par CFE.
Chaque série-année est pourvue d'un numéro de série identique pour toutes les années d'une série et d'un numéro d'année. Celui-ci ne contient que les chiffres significatifs de l'année (68 et non 1968, par exemple). Soit NS le numéro de série et A l'année correspondant à une série - année, le numéro de série - année NSA est donné par NSA = NS* 100 + A.

La structure de l'article est alors la suivante :

a) Cas d'une série sommable :

1er mot	:	le numéro de série-année. Il est enregistré sous la forme "image-miroir".
2° mot	:	Code de facteur d'échelle.
du 3ème mot au n + 2 mot	:	Les u valeurs rendues entières de la série-année (n = 90 ou 95).
n + 3ème mot	:	numéro de la série-année suivante dans le fichier 2. Ce numéro est égal à 999999 pour la dernière série-année du fichier.

b) Cas d'une série sommable

les n + 2 premiers mots (n = 90 ou 95) sont structurés de la même façon qu'en a).

mot n + 3) mot n + 4)		Ces deux mots entiers ne sont que les 48 bits d'un mot flottant qui représente la valeur réelle correspondante de la série pour le niveau "France Entière".

STRUCTURE de l'ENREGISTREMENT

mot 1 mot 2 _____ n mots _____ mot p mot m

Données p + 1

	Série année							
	Sommable				non sommable			
CFE	100				100			
	FR		FD		FR		FD	
	21	22	90	95	21	22	90	95
p	-		-		24	25	93	98
m	24	25	93	98	26	27	95	100

mot n + 5 mot de chaînage. Il joue le rôle du mot n + 3 du cas a).

2.2. - Le fichier régional (F.R.)

Ce fichier a la même structure que le fichier départemental. Seule la valeur de n change : n peut prendre les valeurs 21 ou 22 suivant que le découpage considéré inclut la Corse dans la Provence-Côte d'Azur.

2.3. - Principe de l'enregistrement

Pour les deux fichiers, chaque article constitue un bloc de n mots qui est écrit sur la bande par l'ordre FORTRAN de sortie asynchrone. Aucune autre information que celle décrite ne figure dans l'enregistrement. Celui-ci est donc réalisé sur le mode binaire pur.

En vue de conserver toute l'information disponible et de réduire l'encombrement sur bandes magnétiques, le SAD a toujours cherché à enregistrer l'information la plus fine et de façon non redondante. C'est pourquoi lorsqu'une série sommable est enregistrée dans le fichier départemental, elle ne figure pas au fichier régional. Celui-ci ne contient donc que les séries non disponibles au niveau départemental et les séries qui, disponibles au niveau départemental, sont non sommables.

Le principe de fonctionnement du système d'enregistrement est le suivant : des programmes dits d'enregistrements traitent les données à entrer dans les fichiers, suivant leur nature. Ces données sont perforées sur cartes. Ces programmes se livrent à des vérifications. Ils génèrent un fichier carte. Celui-ci, trié, est alors fusionné avec les fichiers existants. Le programme de fusion admet les modifications à faire pouvant résulter des vérifications.

3- La mise à jour

3.1. - Le programme 750-A

Ce programme a pour objet la constitution sur cartes du fichier de nouvelles données à fusionner avec les fichiers FD et FR.

Le programme admet en entrée un flot d'informations perforées en format libre d'après un bordereau. Ce bordereau décrit les documents où sont imprimées les série-année qui suivent dans l'ordre du bordereau. Elles sont dotées d'un numéro. Le bordereau prévoit l'inscription d'une valeur appelée ordre. En effet, les valeurs de la série peuvent se présenter sous dix ordres différents. La présentation des données sur les documents est réduite au maximum, la perforation se fait en format libre, comme pour le bordereau ; Celui-ci porte un index, dont le premier caractère est R ou D suivant que les séries-années sont régionales ou départementales. Le programme requiert deux bandes de manoeuvre où sont stockés les numéros de séries par type de fichiers ; il procède à la vérification comptable suivante : la valeur France entière est perforée et calculée d'après les valeurs de la série ; si cette valeur est comprise entre le maximum et le minimum de la série, la série est décrétée non sommable ; dans le cas contraire, la valeur lue est confrontée avec la valeur calculée. En cas d'erreur, un listing de la série complète est sorti. Dans tous les cas le programme crée un fichier de fusion dont la structure est la suivante. L'article représente une série-année et est reporté sur plusieurs cartes. Chaque carte porte dans les colonnes 1 à 8 le numéro de série-année , puis suivent les valeurs par zone de 8 caractères y compris le code facteur d'échelle qui est déterminé par le programme. La première carte porte le numéro de bordereau et le nombre de cartes de l'article.

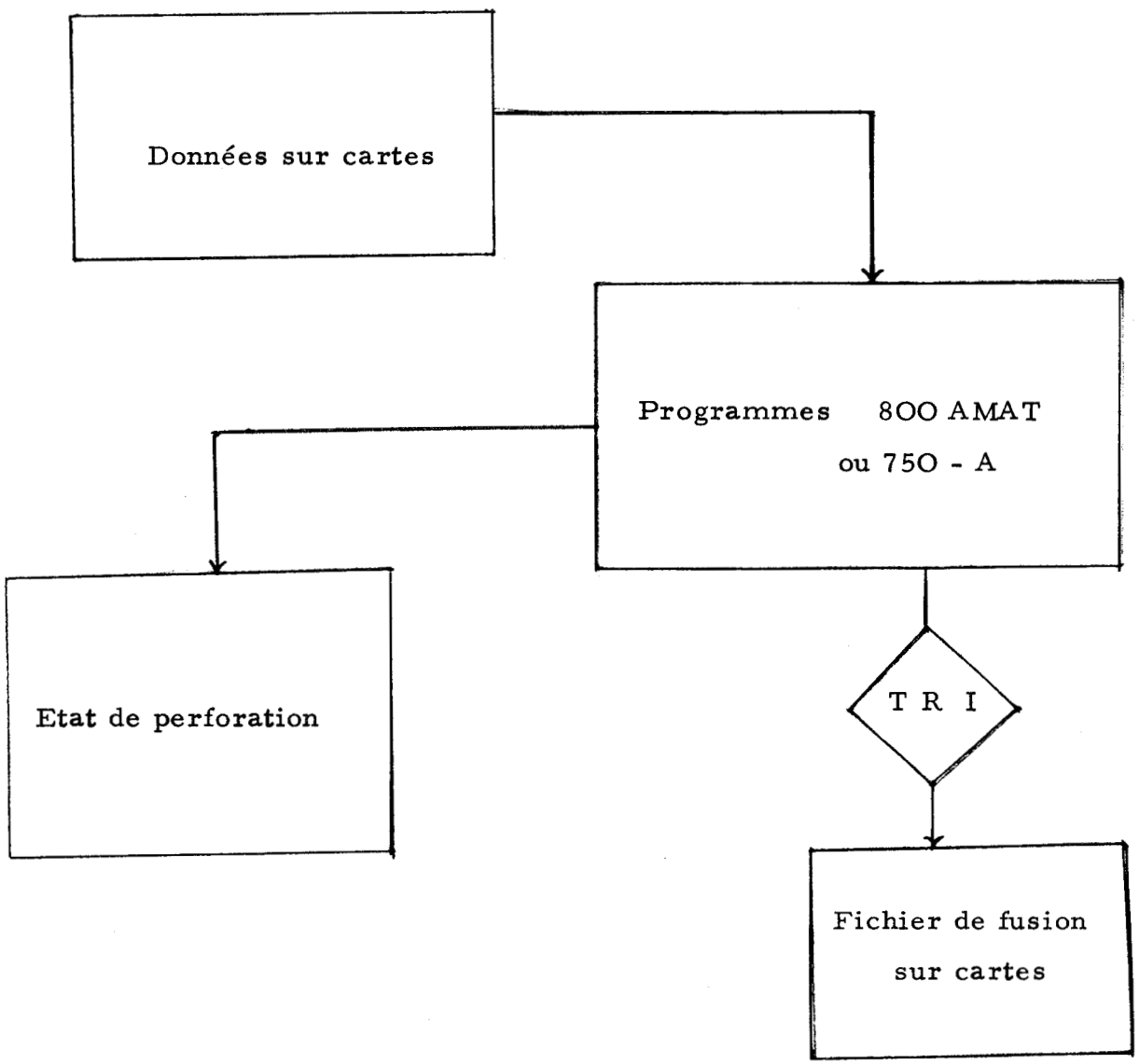
Ce fichier trié sur les numéros de série-année croissants constitue le fichier de fusion.

3.2. - Le programme 800 - AMAT

Ce programme est utilisé lorsque l'information se présente comme un croisement de deux caractéristiques et d'une localisation. Par exemple, la série-année "population active, de sexe masculin, par catégorie socio-professionnelles (à 9 postes) et par tranches d'âge (6 tranches), par département en 68" peut se présenter sous la forme de 95 tableaux à double entrée.

Un bordereau décrit les tableaux avec les lignes et colonnes marginales qui sont perforées le cas échéant. Ceci permet des vérifications croisées et la localisation des erreurs. Le programme génère un état des séries avec erreurs éventuelles et un fichier de fusion.

PRINCIPE de L'ENREGISTREMENT



3.3. - Création des modifications

L'état de perforation est soumis pour inspection aux personnes chargées de l'enregistrement. En cas d'erreur, ces personnes émettent un bordereau de modifications des série-année érronnées. Les modifications sont de plusieurs types.

Type S : suppression de la série-année,
Type M : modification de la valeur d'une localisation,
Type E : modification du facteur d'échelle,
Type T : modification de la valeur "France entière" pour série-année,
Type O1 : transformation d'une série-année sommable en non sommable avec introduction de la valeur "France entière"
Type O2 : transformation inverse.

Ces modifications peuvent également affecter des série-année sur bande et sont également portées sur bordereau de modifications.

3.4. - Le programme 75O-B.

Ce programme assure la gestion du fichier départemental et du fichier régional. Pour un type de fichier donné, il assure la fusion du fichier carte du programme 75O-A, et procède aux modifications du fichier en cours.

Il admet en entrée :

1 - le fichier départemental ou régional, sur bande magnétique,

2 - un fichier sur carte éventuellement vide de modifications de série-année figurant au fichier en cours,

3 - un fichier éventuellement vide de modifications des ajouts sur carte,

4 - le fichier des ajouts issus du programme 75O-A. Ce fichier sur carte peut être éventuellement vide.

Il admet en sortie :

1) le nouveau fichier départemental ou régional sur bande magnétique

2) un état de fusion

Cet état comporte :

a) les valeurs des série-année ajoutées

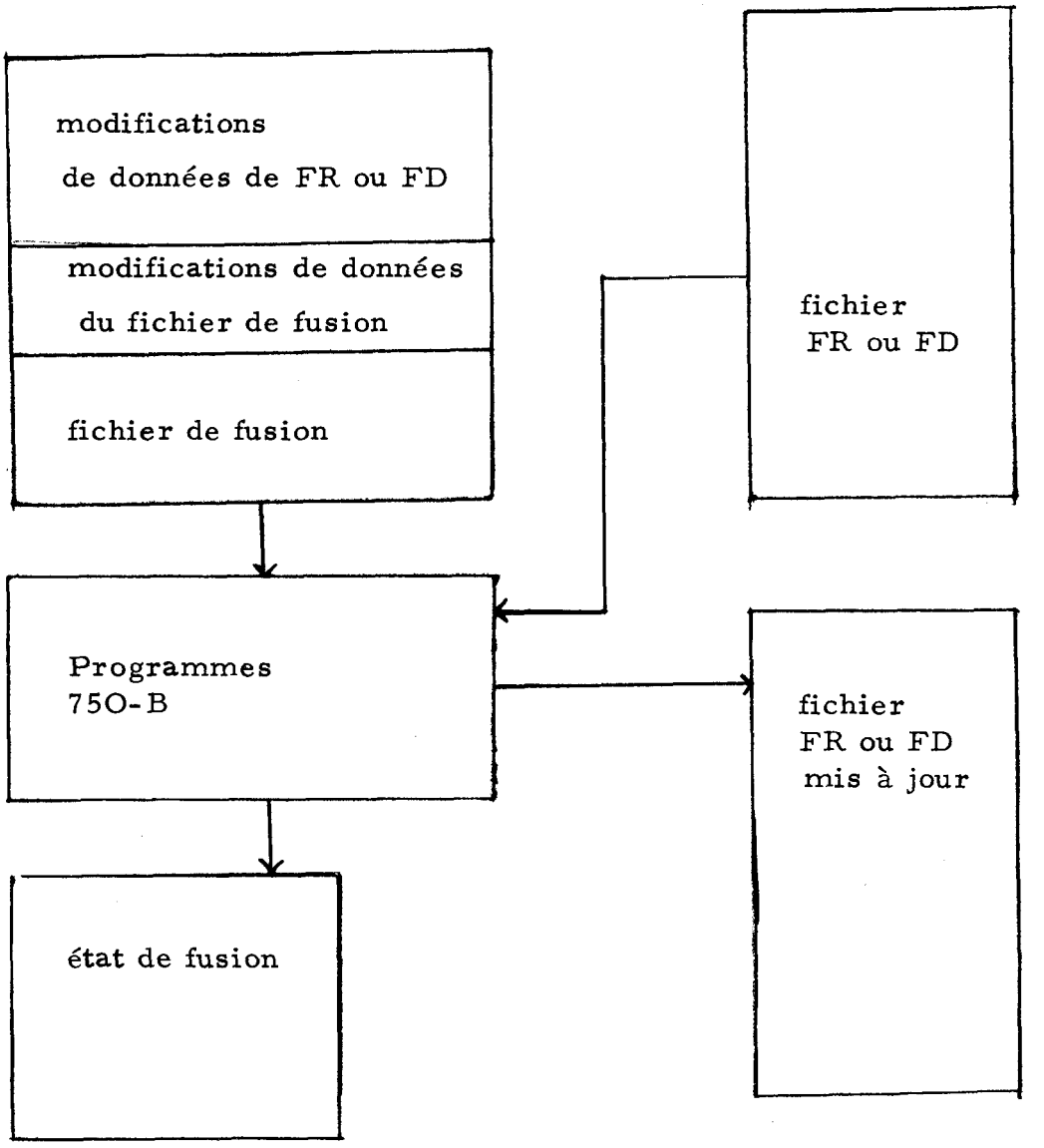
b) les valeurs de l'ancienne série-année lorsqu'elle a été supprimée du fichier

c) les modifications (anciennes et nouvelles valeurs)

d) la liste de tous les numéros de série-année du fichier.

Ce programme permet d'effectuer 300 modifications de série-année dans les fichiers et autant dans le fichier d'ajouts. Il permet la fusion d'un nombre limité à 9999 de série-année.

PRINCIPE DE LA MISE A JOUR



3.5. - Le programme B - ROSELYNE

Ce programme a pour but de sélectionner sur deux fichiers départementaux (ou régionaux) différents des série-année et d'en faire un nouveau fichier départemental ou régional. Les fichiers mis en oeuvre sont tous sur bandes magnétiques, hormis les numéros de série-année sélectionnées qui sont sur carte.

3.6. - Fonctionnement de la mise à jour

L'ensemble des programmes précités est mis en oeuvre facilement par des économistes chargés de l'enregistrement des données. Ces programmes assurent l'interface entre des fichiers de données sur support imprimé et le système SAD I. C'était, alors que peu d'organismes collecteurs d'informations possédaient des ordinateurs, le meilleur moyen d'assurer l'enrichissement de la base de données de SAD I. Cependant, au fur et à mesure que l'équipement informatique des administrations et organismes divers s'accroissait, SAD I devait résoudre des problèmes d'échanges d'informations entre base de données dans les meilleures conditions. Pour cela il fallait éviter les transcriptions nouvelles et les perforations autant que faire se pouvait. Aussi SAD I esseya-t-il de mettre en oeuvre sur des cas d'expériences des programmes permettant de lire des bandes magnétiques d'origines diverses et d'extraire des fichiers qui y étaient enregistrés les informations sous format SAD pour les adjoindre à sa base de données. Il s'agit d'une première approche du problème d'interconnexion de banques de données. Les conclusions qu'on peut en tirer sont qu'en vu de la multiplicité des matériels et des structures de fichiers, il serait bon d'engager des études sur les modèles de fichiers fixant les paramêtres hardware et software permettant l'échange entre systèmes divers implantés sur différents matériels.

4 - Le traitement numérique

SAD I met à la disposition des utilisateurs un ensemble de moyens permettant le traitement des données statistiques. D'autre part, la simplicité de structure des fichiers permet leur utilisation par des programmes utilisateurs en plus du programme principal de la banque de données.

Ce programme permet aux utilisateurs d'utiliser un cadre standard pour l'obtention de tableaux dont les lignes sont indicées par les localisations choisies, dont le nombre de colonnes est limité à 15, et dont le contenu d'une case est obtenu par traitement de série-année de la base de données. Ce traitement comprend des calculs arithmétiques et des éditions surimprimantes. Il est obtenu par utilisation d'un langage simple et relativement rigide.

4.1. - le langage

L'utilisateur a la possibilité de mener ses traitements dans une zone de mémoire centrale accessible par le langage. Cette zone a la structure suivante : 10 colonnes et 11 lignes représentant les départements et les régions. L'utilisateur, en fait, ne connait que les colonnes sous lesquelles il introduit des données départementales ou régionales.

Le langage comporte 4 genres d'instructions, :

les instructions - titres
les instructions - arithmétiques
les instructions d'impression
les instructions de contrôle.

Chaque instruction s'écrit avec l'alphabet FØRTRAN sur 80 caractères au plus.

<u>L'instruction titre</u>

Le type d'instruction est le symbole ✕. L'instruction est terminée

par une chaîne de 79 caractères au plus choisis parmi ceux de l'alphabet FØRTRAN classique.

L'effet de cette instruction est de provoquer l'impression surimprimante de la chaîne de lettres qui suit le type d'instruction.

Les instructions arithmétiques

Elles ont la structure suivante :

⟨type d'instruction⟩⟨operande⟩⟨opérateur⟩⟨operande⟩ ⟨résultat⟩

Le type d'instruction est le caractère 2.

Un operande est soit un operande scalaire, soit un operande-tableau.

Un operande-tableau est constitué par un type de tableau suivi d'une référence de tableau, suivi d'une référence d'année. Une référence d'année est constituée par un numéro d'année simple ou indexé . Un numéro d'année simple est constitué par un entier sans signe de 2 chiffres. Un numéro d'année indexé est composé d'un numéro d'année simple et d'une variable d'index (identificateur I ou J). Une référence de tableau est soit un numéro de série, soit une référence de colonne. Un numéro de série est soit un numéro de série simple (entier sans signe à 4 chiffres), soit un numéro de série indexé qui est un numéro de série simple suivi d'un index (identificateur I ou J). Une référence de colonne est un identificateur de colonne (caractère X) suivi d'un numéro de colonne qui peut être simple ou indexé.

Exemples d'operande

- scalaire	N		. .	1OO
- Tableau	R 1226		63	
	D 1341		I 64	
	D 1341		I 64	J
	R 1226		63	J
	D	X	3	J
	R	X	4	

un opérateur est l'un des caractères suivant :

=, +, - X, / ° = est un opérateur unaire, les autres sont binaires.

Un résultat est un type de tableau (R ou D) suivi d'une référence colonne.

Le type de tableau correspond à la nature de l'information traitée (R = régional, D = départemental, N = scalaire égal à tous les éléments d'une série départementale/régionale).

Une référence de tableau sert à manipuler les données. Une référence de colonnes repère une colonne de la zone de travail ; son numéro est égal soit au numéro de colonne soit au numéro de colonne additionné de la valeur de l'index s'il existe X 1 représente la colonne 1 , X 1 J représente la colonne 1O si J est porté à 9 .

Un numéro de série est la clé d'accès aux données du fichier départemental ou régional. Le numéro de série est égal au numéro de série simple additionné s'il y a lieu de la valeur de l'index. Il en est de même pour les numéros d'années. L'opérateur = est l'opérateur de transfert correspondant à une affectation du contenu de l'operande au résultat. Les autres ont les significations habituelles des langages de programmations tel FØRTRAN ou ALGOL.

Les variables d'indexations sont contrôlées par les instructions de controle.

Instructions de controle

Elles ont la structure suivante :

⟨type d'instruction⟩ ⟨ portée⟩ ⟨ partie indexation⟩

⟨partie indexation⟩ : : =⟨instruction d'indexation⟩ |

⟨instruction d'indexation⟩ ; ⟨ instruction d'indexation⟩

⟨ instruction d'indexation⟩ : : =⟨variable d'indexation⟩ =(⟨liste d'indexation⟩)

<liste d'indexation> : :=<indexation> ; <liste d'indexation><indexation>

<indexation> : :=<entier sans signe> < liste >

< liste> : :=<entier sans signe> ; <entier sans signe> |

<entier sans signe>, <entier sans signe>, <entier>

Exemples d'instruction de contrôle :

1 3 I = (1 2 3 5 ; 1 2 4 1 , 1 2 3 8 , - 1 ; 7040 ; 7045)

1 3 J=(6 5 , 6 7 ; 6 9)

1 3 I = (1 2 3 5 , 1 2 4 1 , 2) ; J = (5 3 ; 5 5 , 6 7,2)

Le type d'instruction est 1. La portée représente le nombre d'instructions qui suivent en séquence et sur lesquelles porte l'indexation. Celles-ci donnent des valeurs aux index I et J qui les prennent dans des listes de valeurs définies par celles-ci, soit par une borne supérieure et une borne inférieure avec un pas facultatif lorsqu'il vaut 1, soit par une suite d'entiers.

L'index I sert à indexer les numéros séries; tandis que J est utilisé pour l'indexation des années ; tous les deux peuvent être utilisés en simultanéité. Il en est de même pour l'indexation des colonnes.

- Instructions d'impression.

Leur structure est la suivante :

<type d'instruction><liste de sortie> ;<liste de formats > |

<type d'instruction><liste de formats> ;<liste de sortie><liste de sortie>

<liste de sortie>. :=<type de tableau><liste de valeurs>

<type de tableau>: :=R | G | D

<Liste de valeurs>: :=<liste de valeurs> , <valeur> | valeur

<valeur : :=<entier sans signe> | (<entier sans signe> <entier sans signe>) |

(<entier sans signe> , <entier sans signe> , <entier>)

<liste de format>: :=<format> | <format> ; <liste de format>

<format>: :=liste * P<facteur de cadrage> V<entier sans signe> -

<liste de valeurs>

<liste *> : : = * | ** | ***

<type d'instruction> : : = 3

facteur de cadrage : : O | 3 | 6

Le type de tableau provoque une addition départementales, régionales ou départementales-régionales (cas des départements groupés en régions) suivant sa valeur D, R ou G.

La liste des valeurs s'interprète comme une liste de numéros de colonnes de la zone de travail sur lesquelles vont porter les éditions. La forme parenthèsée est une forme avec bornes inférieure et supérieure et pas facultatif s'il vaut 1. La liste de sortie provoque l'impression des colonnes spécifiées dans un format standard (I10). Celui-ci peut être modifié par la liste de format. Cette liste est limitée à 3 formats. Chaque format comporte un facteur de cadrage égal à O, 3, 6. L'enregistrement sortie est égal à la valeur réelle multipliée par 10 puissance moins le facteur de cadrage (sorties en unités milliers, millions). Le nombre de chiffres après la virgule figure après la spécification V. Un format porte sur un groupe de colonnes référenciées par la liste de valeur qui les suit. Chaque format est précédé d'une liste * qui ne peut prendre que 3 aspects.

exemple d'instructions d'impression :

3 D (1, 10)

3 R (7, 13, 2) , 1 , 2 ; * P 3 V1 - (7,13) ; ** OV3 - 1***P6 V2 - 2

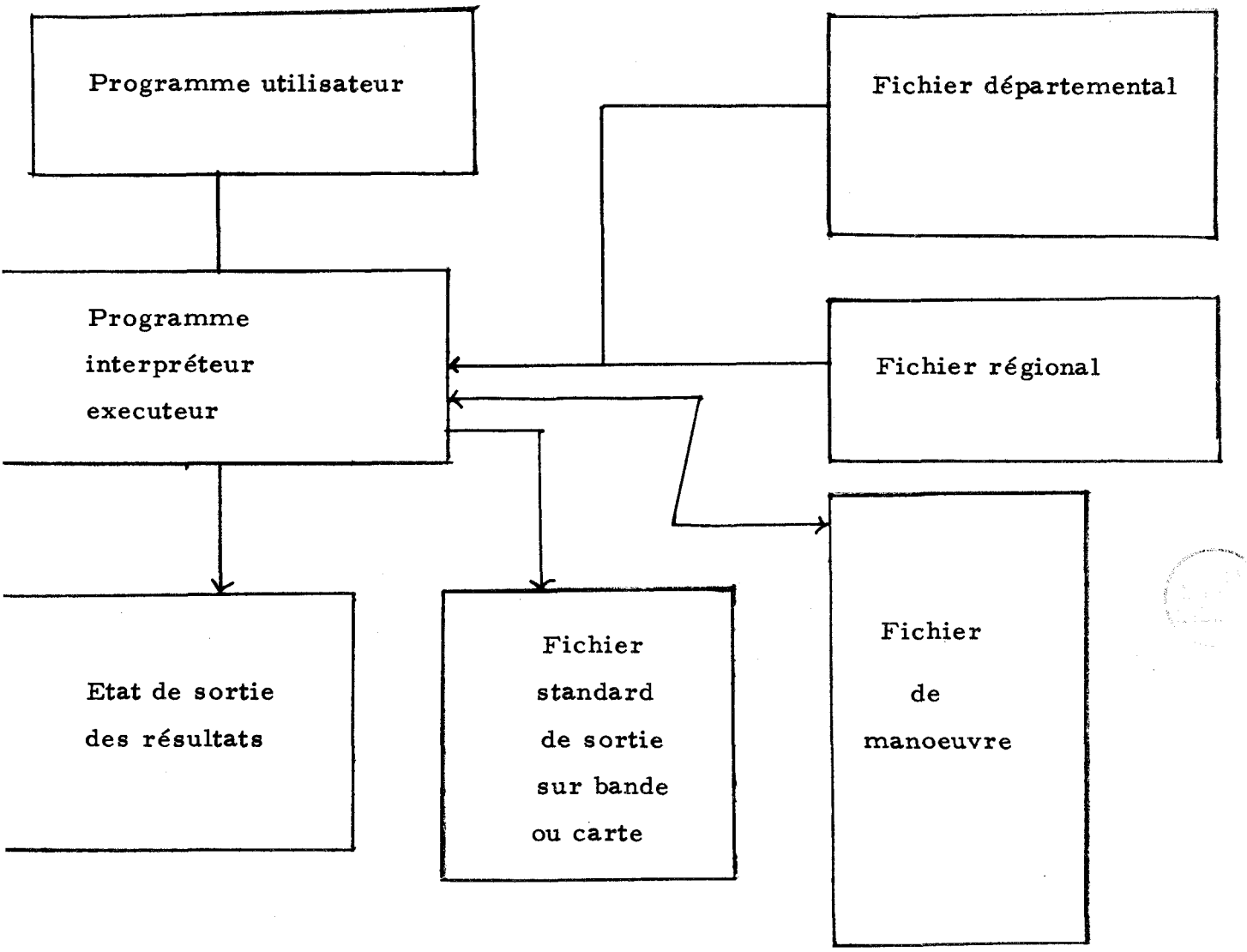
Ce langage permet donc un traitement des informations contenues dans les fichiers départementaux. Ces traitements s'appliquent toujours à des séries localisées et l'impression ne peut être provoquée que sur trois modes : le mode départemental (D) , le mode régional (R), le mode départemental - régional (G).

4.2. - Le programme 750-CORSE

Le programme d'exploitation de la banque de données se présente comme un interpréteur du langage de traitement.

Les instructions sont perforées, soit dans un format fixe qui dépend du type d'instruction, soit dans un format libre pour certains types d'instructions. Elles constituent le fichier carte en entrée du programme.

PRINCIPE DU TRAITEMENT NUMERIQUE



Ce programme se déroule en plusieurs "phases" appelées passe et nécessite une mémoire de 32 K mots et 6 dérouleurs de bandes magnétiques. Il est écrit pour partie en SYMBOL (Assembleur du CAE 90-80), pour partie en FORTRAN IV CAE.

La passe O consiste en une écriture sur bande magnétique de toutes les instructions, après vérification syntaxique. Le programme procède également à la création d'une table des numéros de série-année contenus dans les instructions. Cette table est triée, débarrassée des articles en double, puis écrite sur bande magnétique. Elle est utilisée dans le cas de série-année demandées qui n'existeraient pas. Le programme extrait des fichiers départemental et régional les deux sous-fichiers utiles, composés des série-année demandées. La passe O se termine par un rembobinage général.

La passe 1 consiste en une lecture et interprétation partielle des instructions sur bande et la confection de la table initiale de la passe O. Avec cette table, le programme à partir des deux sous-fichiers crée deux nouveaux sous-fichiers où les articles (série-année) sont placés dans l'ordre de leur apparition dans les instructions, avec, le cas échéant, des répétitions éventuelles.

Cette passe se termine par un rembobinage général.

La passe 2 explore la bande magnétique où sont écrites les instructions, les interprète et les exécute. Les données sont extraites des 2 sous-fichiers "utiles"

Le programme peut également sortir des résultats sur cartes ou bandes magnétiques hors un format déterminé.

L'utilisation, pas à pas, du programme est possible, avec comme support d'entrée des instructions, la machine à écrire du pupitre de l'ordinateur.

4.3. - Fonctionnement du traitement numérique

La mise en oeuvre de ce programme 750-CORSE n'est pas facile car le nombre de fichiers de manoeuvre est important. L'utilisation de la mémoire centrale est poussée à son maximum. Moins de 150 mots de la mémoire centrale

disponible ne sont pas utilisés. L'introduction de la Corse dans l'ensemble des régions portant leur nombre à 22 ne peut pas être pris en compte dans tous les cas. Seul, le cas où l'information est traitée au niveau départemental aura impression des résultats au niveau régional. Enfin ce programme procède d'une conception séquentielle des fichiers, de même que les programmes de mise à jour. La généralisation des mémoires à accès direct rend quelque peu caduque la conception d'ensemble de ces programmes. Enfin, le traitement "batch" obtenu par le programme 750-CORSE est de nature à allonger les temps de réponses. Une incorrection syntaxique peut empêcher le déroulement correct d'un grand nombre d'instructions. Le langage, quoique fixé dans un cadre rigide d'écriture, exige un apprentissage de la part de l'utilisateur ou l'intervention d'un spécialiste.

5 - La Recherche des Séries par leurs titres.

On appelera ici titre le descriptif d'une série. Pour un ensemble de série-année d'une même série, les descriptifs ne différent des titres que par l'adjonction d'un terme de localisation dans le temps (date). Le problème abordé ici est le suivant : l'utilisateur désire connaître les données disponibles dans le système et qui peuvent servir à traiter son problème ; une variante du problème est que l'utilisateur, sachant que le système peut lui fournir une donnée, cherche à savoir si cette information n'est pas disponible avec une plus grande finesse.

Le système doit donc recevoir des demandes de titres de séries de la part de l'utilisateur et lui fournir, en réponse, les titres qui satisfont sa demande.

Deux problèmes se posaient :

1) le codage de l'ensemble des titres devait se faire de la façon la plus économique qui soit et devait permettre un temps de réponse acceptable, compte tenu de la configuration de l'ordinateur employé (BGE M.40).

2) Les réponses devaient être faites sans un trop grand "bruit" mais le système devait compter avec cependant une certaine latitude dans la rédaction de la demande.

Les problèmes sont, grosso modo, ceux posés en partie par la documentation automatique. On aurait pu, puisque l'objet traité est sensiblement de même nature (équivalence d'un titre de série avec un abstract), utiliser un système déjà existant. Cela aurait conduit à méconnaître la spécificité du problème qui tient au contenu sémantique de la référence, et à la structure combinatoire de l'ensemble des titres. Aussi a-t-on tenté d'élaborer un langage d'analyse des titres qui puisse fournir une structuration de l'ensemble et un codage pertinent.

Enfin, le système a été monté sur l'ordinateur BGE M4O du laboratoire de calcul de la Faculté des Sciences de Lille. Cet ordinateur avait une mémoire de 16 K mots de 24 bits et possédait les périphériques suivants :

- 1 tambour de 32 K mots,
- 4 dérouleurs de bandes dont 3 disponibles pour l'utilisateur,
- une imprimante,
- un lecteur de cartes,
- une machine à écrire.

Parmi les langages disponibles, on a choisi l'ALGOL pour l'écriture du programme, langage symbolique qui sur le M4O est d'un emploi plus commode que le FØRTRAN.

5.1. Le langage de description des données statistiques.

Les données statistiques résultant soit d'une ventilation des éléments d'un ensemble par rapport aux caractères de certaines caractéristiques, soit d'une totalisation d'un critère sur les éléments également ventilés. Le titre d'une série reflète donc d'abord la nature de l'opération faite pour l'obtenir, ensuite précise les caractères de croisements dont l'un est particulièrement important. Il s'agit du caractère de localisation. Dans une répartition par région, chaque région est un caractère de la caractéristique "région". Le SAD traite donc deux caractéristiques de localisation : les départements, les régions.

Le titre d'une série est en fait le nom d'une opération faite sur un sous-ensemble obtenu par division d'un ensemble donné. Si, à tout sous-ensemble obtenu par division, on associe un nom, celui du caractère de la caractéristique associée à la division, l'énumération des noms des sous-ensembles successifs résultant des divisions successives permet de caractériser le sous-ensemble attaché à la série. La méthode des divisions, grâce à la suite des caractères des caractéristiques associées, est un moyen de caractériser un sous-ensemble de l'ensemble de départ. Ce dernier correspond dans le langage à un terme d'élément. Aux sous-ensembles d'une division sont associés les termes de caractères, à la division, elle-même, est associé le terme de caractéristique.

Considérons le titre suivant :

"Nombre de personnes, actives, de sexe masculin, célibataires, âgées de 35 à 45 ans, par département".

Le terme "nombre de" caractérise l'obtention par ventilation d'une population en sens statistique du mot. Le terme d'élément est ici "personnes". Il s'agit d'une division de l'ensemble de personnes qui a permis la ventilation. La première caractéristique employée est la distinction "par activité", terme de caractéristique, dont le caractère retenu est ici "actives", en opposition à l'autre caractère de cette caractéristique qui est "non actives". La deuxième caractéristique est le sexe, dont le terme s'écrit "par sexe". On a retenu le caractère de "sexe masculin"; la troisième est "par type de statuts", le caractère étant ici "célibataires"; enfin, la dernière caractéristique, "par tranche d'âge" termine la description par le choix du caractère de "âgées de 35 à 45 ans". Le terme de localisation "par département" indique que la série est départementale. Il s'agit d'un terme de caractéristique. Le terme "nombre de" est appelé terme général. On voit qu'on engendre un titre de série par la concaténation du terme général, du terme d'élément, d'une chaîne de termes de caractères. Si dans cette chaîne, à un caractère, on substitue le terme de la caractéristique à laquelle appartient le caractère, on engendre autant de titre de séries qu'il y a de caractères dans la caractéristique. On voit ici apparaître le caractère combinatoire. Si donc, la caractéristique "par tranche d'âge" est à 9 caractères, si la caractéristique par "type de statut" est à 4 caractères, la caractéristique "par activité" à deux caractères, le titre suivant :
"nombre de personnes, par activité, par sexe, par type de statuts, par tranche d'âge, par département"
représente en fait 2 x 2 x 4 x 9 = 144 titres de séries. On construit ainsi de façon économique, par combinaison de termes, des titres en indiquant la seule référence aux caractéristiques et un titre particulier en indiquant la référence des caractères des caractéristiques employées.

Caractéristiques et caractères localisent un sous-ensemble dans un ensemble

donné. Une autre manière de le faire est d'établir une relation entre deux sous-ensembles différents. Si la relation est "vraie", le sous-ensemble est parfaitement déterminé. Par exemple, soit le titre suivant :
"nombre de personnes, par sexe, salariées des entreprises, par catégorie d'activité économique".
Le sous-ensemble des personnes, déterminé l'un des caractères de la caractéristique "par sexe" et le caractère "salariées" est également caractérisé par la relation "des". Ce terme explicite une relation entre un sous-ensemble des "personnes" et un sous-ensemble des "entreprises". Ainsi, la "relation" entre éléments caractérisé est une caractérisation de l'élément "gauche". Notant E un terme d'élément, C un terme de caractéristique ou de caractère, on peut construire un terme d'élément caractérisé sous la forme $E(C^{*}\ REC^{*})^{*}$ où * représente la répétition de la chaîne à laquelle il s'applique.

Une série peut aussi être la totalisation d'un critère attribué à un sous-ensemble déterminé ou encore une opération sur des totalisations. Ainsi on aura "montant des impôts des salariés des entreprises, etc", ou encore "superficie des terres labourables, cultivées en blé dur", ou encore "densité des personnes âgées de 5 à 15 ans". L'énoncé de la totalisation constitue un terme de valeur. Le titre commence alors par un terme de valeur et un terme d'unité facultatif. Ces termes sont peu nombreux.

On peut donc énoncer les règles de production des titres de séries :

- ⟨titre⟩ : :=⟨début de titre⟩ ⟨terme d'élément caractérisé⟩ ⟨terme de localisation⟩

- ⟨début de titre⟩ : :=⟨terme général⟩ | ⟨terme de valeur⟩ | ⟨terme de valeur⟩ ⟨terme d'unités⟩

- ⟨terme de localisation⟩ : := "par région" | " par département"

- ⟨terme d'élément caractérisé⟩ : :=⟨terme d'élément⟩ | ⟨terme d'élément caractérisé⟩ |

<terme de relation><terme d'élément caractérisé>

|<terme d'élément><liste de caractérisation>

- <liste de caractérisation> ::= <terme de caractérisation> |

<terme de caractérisation><liste de caractérisation>

-<terme de caractérisation>: :=<terme de caractère> <terme de caractéristique>

-<terme de caractéristique>: :="par sexe" | "par destination"|...

-<terme de caractère>: :="de sexe masculin" | "à toit de zinc"|....

-<terme d'élément>: :="personnes"|"entreprises"|....

-<terme de valeur>: :="superficie" | "valeur" |...

-<terme d'unité>: :="en hectare"|"en franc"|"en millier de kilomètres"|..

La notion de titre introduite représente alors un ensemble de titres de série obtenu en faisant varier les caractères des caractéristiques ou en les fixant. Le vocabulaire est emprunté au français. La grammaire engendre ainsi des phrases françaises.

Ce langage sert à analyser des documents qui supportent des données. On dresse ainsi le "catalogue des données" en rédigeant en LDDS les énoncés des titres. Le vocabulaire est créé par l'analyste. Ce vocabulaire n'est pas très étendu et permet de décrire un très grand nombre de données. Ainsi, moins de 200 termes permettent de décrire les 14 000 séries de recensement industriel.

Pour exploiter ce langage, il reste à coder les termes et les titres. On gagne ainsi en mémoire une place très importante en ne stockant que le vocabulaire.

Le codage est assuré par l'analyste au moment de la description des données.

Le langage créé alors restait insuffisant mais il permit de tenter une expérience d'analyse de quelques 17000 titres de série et de créer un programme de recherche de séries sur ces bases. Le développement des idées de base de LDDS conduiront dans SAD II à des conclusions plus riches et notamment permettront de résoudre quelques problèmes de "bruit".

5.2. - Principe de la recherche - Mise en oeuvre.

Le codage des termes s'impose pour tenter une représentation simple et économique des titres. Tout terme est codé sous forme d'un code alphabétique et d'un entier sans signe.

On a représenté :

terme général	par	X (l)
terme d'élément	par	EL (x)
terme de valeur	par	U (x)
terme d'unité	par	V (x)
terme de relation	par	R (x)
terme de localisation	par	LA (x)
terme de caractéristique	par	α (O)
terme de caractère	par	α (x)

α représente une chaîne de caractères, et x un entier sans signe.

La représentation d'un ensemble de titres obtenus en faisant varier le caractère d'une caractéristique ou de plusieurs, nécessite un codage de la liste des caractères utilisés pour une caractéristique. Le terme de caractéristique équivaut à la liste complète de tous les caractères de celle-ci. Une liste de caractères pour la caractéristique codée α peut se définir soit avec une borne inférieure, une borne supérieure et un pas, soit comme une liste simple, soit encore comme une liste composée de listes conformes aux types précédents.

Le premier conduit à la représentation x, y, p, le second à la représentation u ; v ; w où x, y, p, u, v, w, sont des entiers sans signe. La liste composée utilisera le caractère ; comme séparateur. Une liste de caractères d'une caractéristique s'écrit :

α(x) ou α(x, y, p), α(u ; x) ou encore α(u ; x, y, p;v;....)

On peut ainsi représenter des ensembles de titres de façon très concise.

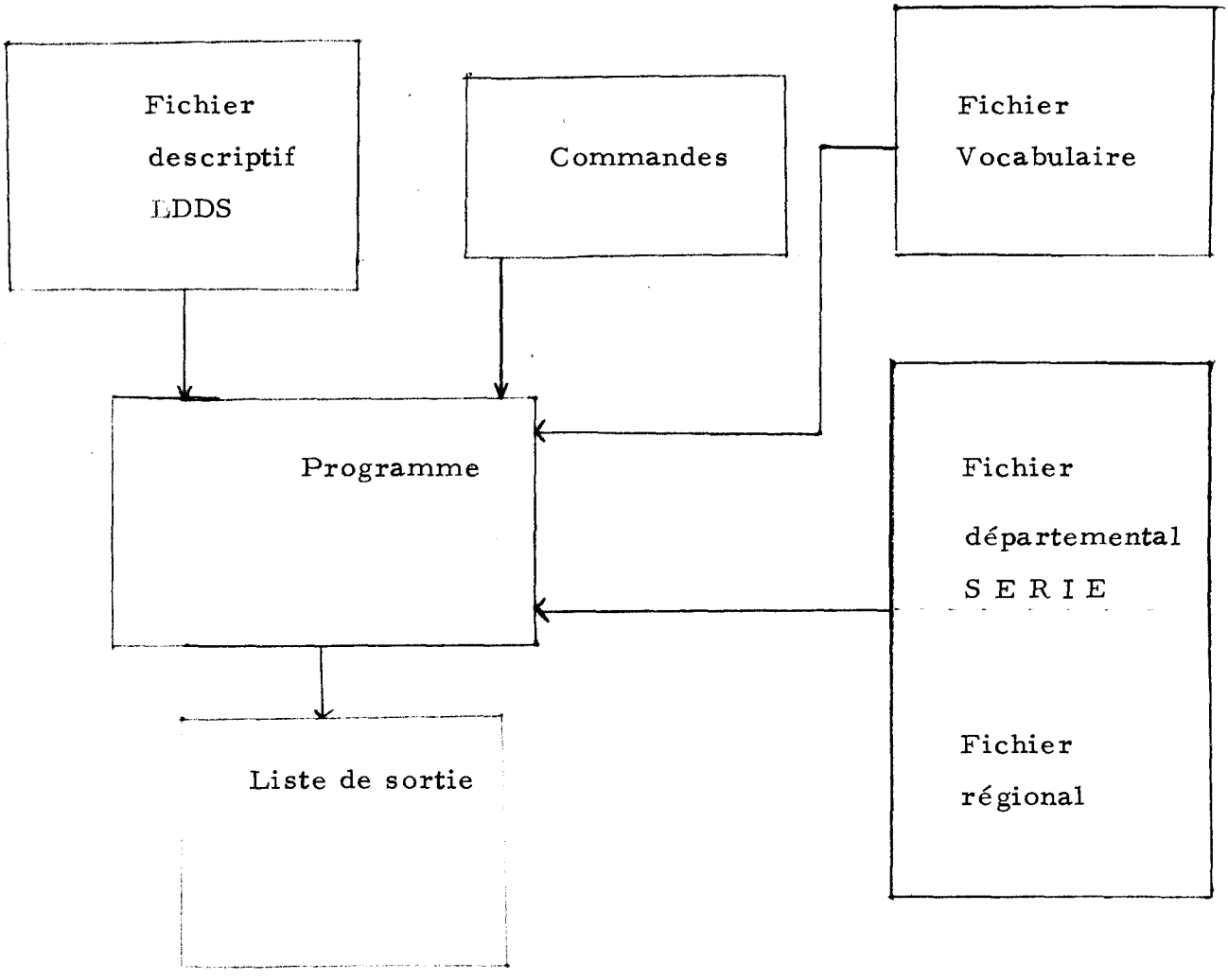
Par exemple, "nombre des entreprises ; par type d'entreprise, par taille d'entreprise en nombre de salariés, par secteur d'activité" équivaut à : 4 x 3 x 161 = 1972 titres élémentaires de séries ; car il y a 4 types d'entreprises, 3 tailles, et 161 secteurs. Ce titre est codé par :
x (1) EL (1O) RC (1, 4, 1) RE (1, 3, 1) RA (O)
ou par :
x (1) EL (1O) RC (O) RE (O) RA (O)
pour représenter les titres élémentaires correspondant aux secteurs 1 à 1O dans la liste des secteurs, au 15ème secteur et au 47ème pour les formes d'entreprises numéro 3 et 4, on écrit :
x (1) EL (1O) RC (3;4) RE (O) RA (1 , 1O, 1 ; 15 ; 47).

La création des codes et du vocabulaire a été laissée à la charge d'analystes. Le vocabulaire a été organisé par termes de caractéristique et de caractère, termes de relation d'une part, terme général, termes de valeur et d'unité, termes d'élément , termes de localisation d'autre part.

On a appelé domaine un ensemble de données relatives à une source d'information et publiée dans un document précis. L'expérience fut menée sur 4 domaines : domaine du "logement"
domaine du "commerce extérieur"
domaine de la "consommation d'énergie"
domaine du "recensement industriel".

A chaque domaine correspond un ensemble déterminé de caractéristiques et caractères, alors que les termes d'éléments ne sont pas spécifiques d'un domaine tout comme les termes non liés aux éléments ou caractéristiques. Aussi, le "corpus des termes" fut-il divisé en 5 parties. Aux 4 domaines correspondaient

PRINCIPE de FONCTIONNEMENT de la RECHERCHE de SERIES par leur TITRE.



4 parties rassemblant les caractéristiques utilisées dans le domaine, alors que la cinquième rassemblait tous les termes non liés aux domaines. Ce corpus donnera naissance au fichier "VOCABULAIRE" divisé en 5 sous-fichiers.

A chaque domaine correspond un ensemble de titres codés appelés "formules". Le fichier "TITRES" qui rassemble les "formules", représentations codées des titres, sera divisé en 4 sous-fichiers correspondant aux domaines.

Le programme de recherche assure deux fonctions :

- recherche des titres répondant à une demande,
- édition totale ou partielle des titres d'un domaine donné,

qui peuvent s'accompagner, en option facultative, de l'édition des séries dont les titres ont été produits.

Les demandes sont introduites sur cartes ou à la machine à écrire.

A) Demande d'édition :

1 - totale.

On tape l'enregistrement portant en-tête le mot REED.

L'enregistrement suivant est le nom attribué au domaine codé sur 8 caractères à choisir parmi "LOGEMENT"
"REC INDUS"
"CE REN"
"COMM EXTE"

2 - Partielle

L'enregistrement de commande est constitué par :

REED PART X Y

X et Y représentent sur 5 caractères les numéros de titres de

séries dans le catalogue général du domaine considéré, dont on indique le nom bomme dans le cas précédent.

B) Demande de recherche de titres :

L'enregistrement de demande est le suivant :
REC M
L'enregistrement qui suit est une formule codée en LDDS qui représente la question.

Le programme procède à l'impression des titres demandés quand ils existent. Il peut aussi, à la demande, faire une recherche des séries correspondantes dans les fichiers départemental ou régional. Ceci est obtenu en ajoutant le mot de commande TABL aux commandes de recherches ou d'édition. Les série-année correspondant aux séries dont les titres sont la réponse de la demande sont alors imprimées sur la liste de sortie.

L'utilisateur doit donc coder en LDDS sa demande. Pour ce faire, il dispose d'un manuel qui se présente sous la forme d'un dictionnaire avec en entrée un ensemble restreint de termes usuels du vocabulaire français. Chaque terme renvoie soit à un code soit à un autre terme, de telle sorte qu'à la fin de la consultation de ce "livre de codes", l'utilisateur ait pu composer sa demande en code.

5.3. - Structure des fichiers.

Les fichiers sont au nombre de 4. Deux d'entre eux sont les fichiers de séries FD et FR (cf. 4.3.) Les deux autres sont propres au programme de recherche.

- Fichier des descriptifs.

Ce fichier sur bande magnétique contient 4 sous-fichiers correspondant aux domaines. Ils sont constitués par des enregistrements binaires en longueur variable (bloc de n mots). Un sous-fichier s'ouvre sur un bloc de trois mots qui l'identifie et le caractérise, dont la structure est la suivante :

ler mot	:	nombre d'identificateurs de termes utilisés pour le codage du domaine et nombre d'articles spécifiques du fichier. Ces deux nombres écrits en décimal sont concaténés, le premier occupant les 2 premiers chiffres décimaux.
2ème au 3ème mot	:	Les 8 caractères du nom du domaine.

Le deuxième enregistrement est une table des identificateurs alphabétiques de termes avec pour chacun d'entre eux le nombre de termes de la liste qu'ils représentent (cas des caractéristiques et des caractères par exemple).

Les articles spécifiques du fichier représentent les formules codées. Elles sont représentées par des tableaux ayant 6 lignes et un nombre de colonnes variables. Les identificateurs de termes de la formule indicent les colonnes. Pour une colonne, 5 lignes sont réparties pour représenter sous forme entière, 2 listes de numéros représentées par bornes et pas et 3 listes simples. La première ligne contient les identificateurs de terme.

Par exemple, soit la formule codée :

X (1) EL (4) CC (1, 9, 1) CA (1 ; 2 ; 18)

Le tableau qui la représente est :

X	EL	CC	CA
O	O	O10901	O
O	O	O	O
O	O	O	O
O		O	O
OOOOO 1	OOOOO4	O	O10218

L'article est composé d'un enregistrement de 2 mots donnant la longueur (nombre de données) de l'enregistrement suivant, le nombre de titres représentés, et de l'enregistrement du tableau représentant la formule.

Chaque sous-fichier est terminé par une marque de fin de fichier qui est doublée pour le dernier, terminant ainsi le fichier des descriptifs.

- Fichier "vocabulaire".

Le fichier est composé de 4 sous-fichiers contenant le vocabulaire spécifique des 4 domaines et un sous-fichier contenant le vocabulaire général. Le vocabulaire est enregistré par liste de termes. Un article correspond à une liste et comprend deux blocs.

1er bloc

Ce bloc de 3 mots contient l'identificateur de la liste de terme, le nombre de termes et la longueur de l'enregistrement des libellés de termes.

2ème bloc.

Ce bloc dont la dimension est connue par le précédent contient pour tous les termes la longueur en caractères de la représentation du libellé des termes puis les libellés proprement dits, à raison de 4 caractères par mot.

5.4. - Fonctionnement du programme.

Dans le cas d'une recherche, le programme décode la demande et la met sous la forme de représentation des formules codées.

Le programme recherche d'abord le domaine, soit par son nom (cas d'édition de catalogue), soit en consultant la liste des identificateurs des termes du domaine en comparaison avec ceux de la demande.

Une fois le domaine déterminé, l'ensemble des descriptifs est consulté en confrontation avec ceux de la demande. Les tableaux retenus sont alors transférés sur le tambour. Le vocabulaire du domaine est également transféré du fichier "Vocabulaire" au tambour après sélection et restructuration. Un groupe de réponses à une demande est alors constitué par un tableau représentant une formule. L'ensemble des groupes a été empilé sur le tambour. Chaque tableau est alors repris et confronté par son contenu dans sa totalité avec celui qui correspond à la demande. Le tableau final des combinaisons de termes est constitué en mémoire centrale et l'édition des titres a lieu par consultation du sous-fichier "Vocabulaire" stocké sur le tambour. Dans le cas où l'utilisateur a demandé l'édition de la série correspondant au titre qui a été édité en réponse, le programme procède à la recherche de la série dans le fichier correspondant. Ceci est possible, à la suite de la consultation d'une table qui, placée en tête des fichiers de série, est stockée sur le tambour dès le premier accès à ceux-ci. Cette table établit une correspondance entre le numéro d'ordre du titre de la série et le numéro de la série. Elle

devrait pouvoir être supprimée si le numéro de série pouvait être calculé à partir du descriptif. Ceci est possible, si on repère dans la formule codée le titre d'une série par son numéro dans la liste obtenue par le "développement" de la formule. Ce développement est le résultat de la combinaison de tous les termes à prendre dans chaque liste. Une convention de développement identique à celle de l'implantation en mémoire d'un tableau permet donc de calculer le numéro d'un titre correspondant à une combinaison de valeurs prises dans les listes des termes. Cette méthode aurait entraîné une renumérotation complète des fichiers de séries; or, ces fichiers étaient utilisés sur deux ordinateurs différents et dans deux systèmes complètement séparés. Les préoccupations des questionnaires de ces fichiers étaient alors toutes différentes. D'autre part, il aurait fallu procéder un travail de codage beaucoup plus important que ce qui avait été prévu pour cette expérience.

5.5. - Les résultats de l'expérience.

Le programme , pour être mis en oeuvre par un utilisateur, nécessite l'utilisation d'un dictionnaire manuel sous la forme d'un manuel des codes. Il est possible d'automatiser cette phase pour la recherche des titres à partir de la demande en français courant. En effet, il s'agit simplement de repérer les mots-clés dans la demande. Cependant, on n'évitera pas les ambiguïtés du français, car, à un mot-clé retrouvé, correspondent plusieurs termes du LDDS. Une demande en français engendre alors plusieurs demandes codées par un tel système. Beaucoup de ces demandes sont alors éliminées dans le cas manuel, par le jugement de l'utilisateur. Ce jugement s'appuie sur des considérations sémantiques, dont cette version du LDDS ne tient pas compte. Ce point sera amélioré dans SAD II.

La syntaxe d'écriture des titres en LDDS ne joue aucun rôle dans le programme de recherche si ce n'est dans la structuration du vocabulaire puisque une partie du fichier du vocabulaire est composée par les termes ayant une fonction syntaxique particulière (terme d'éléments, général, de valeur, etc ...). une

nouvelle étude, plus approfondie de la syntaxe, permettra dans SAD II de structurer les fichiers descriptifs. Elle permettra aussi de résoudre le problème suivant : une demande de série très précise peut engendrer une réponse très volumineuse, formée par de nombreux titres dans le cas où sont stockées dans le système les séries correspondant à de multiples partitions de la série demandée. La reconstitution par agrégation de séries est possible, souvent de plusieurs manières, si on associe les caractéristiques non plus à des divisions, mais à des partitions. Cette reconstitution devrait pouvoir être reconnue par le système et conduire à une réponse unique.

C'est en s'engageant dans ces voies que l'on pourra ainsi éliminer le "bruit" autour d'une réponse du programme à une demande introduite de façon plus simple par l'utilisateur.

L'utilisation expérimentale du programme a permis de constater que, sur un système à bandes, la méthode mise en oeuvre pour le codage et la recherche, était acceptable du point de vue temps de réponse, si on se limitait aux titres. Les recherches sur bandes magnétiques dans les fichiers régional et départemental restent évidemment très longues. D'autre part, la mémoire centrale utilisable restait relativement limitée et de nombreuses améliorations de la programmation auraient été possibles si la mémoire avait été plus grande. Ce programme devait en effet prévoir le cas de réponses extrêmement volumineuses, qui en mémoire engendrent de grands tableaux. Ce cas a été prévu lors de la constitution du fichier des descriptifs. Chaque représentation d'une formule a été divisée, le cas échéant, de telle sorte qu'elle ne représente pas plus de n titres à p termes tels que n x p soit inférieur à 1500. Cette multiplication des représentations a un effet d'allongement du temps de réponse.

Comme le programme avait été conçu dans le cadre d'une expérience, aucun moyen de mise à jour n'a été prévu.

6 - Caractérisation de SAD I.

SAD I se présente comme un système ayant les caractéristiques suivantes :

- il existe un langage de définition des données. Ce langage n'a pas été utilisé pour la structuration des fichiers de données.

- Ce langage est utilisé pour la recherche de données.

- Il existe un langage traitement des données. Ce langage est interprété.

- Le traitement utilise des méthodes d'accès séquentiels, et s'effectue sur le mode "batch processing".

- La mise à jour s'effectue aussi sur ce mode.

- Le système ne présente que peu d'interfaces avec d'autres systèmes.

- La nature des fichiers de données permet un accès facile à des programmes utilisateurs extérieurs au système.

SAD I est resté un système expérimental en ce qui concerne la partie "recherche de l'information". Opérationnel pour la partie traitement des séries statistiques, SAD I fut exploité assez largement pour qu'on puisse faire quelques remarques sur son utilisation. Le traitement coûte cher en temps machine, du fait de l'accès séquentiel à des fichiers importants. La structure des fichiers ne tient pas compte de la fréquence d'utilisation de certaines données élémentaires et agrégées, si bien qu'il a fallu renoncer au principe de non redondance pour un certain nombre de données. La mise à jour s'est révélée assez facile et assez souple, d'autant plus que l'information économique ne nécessite pas de mises à jour par modification très fréquentes, et que les procédures d'ajouts sont d'un bon rendement. Cependant ces opérations ont été confiées à des spécialistes.

SAD I reste cependant un système difficile d'accès pour l'utilisateur, tout

en permettant des travaux d'une assez grande ampleur. L'expérience acquise avec ce système a permis de concevoir et l'élaborer un nouveau système SAD II qui satisfasse mieux les objectifs assignés à une telle banque de données.

Chapitre II

LE SYSTEME SAD II

Ce système est en cours d'élaboration. Cependant il est possible d'en présenter les grands traits au point de vue structure et fonctionnement. La philosophie qui a pu être dégagée des premières études, a permis la réalisation d'un système léger et transitoire : SAD II TS, qui est décrit plus loin.

2.1 - L'information.

L'information traitée est ici encore double. Il s'agit d'une part des séries statistiques, d'autre part, des descriptifs de séries. Ce descriptif comprend trois éléments, dont seul le premier existe toujours :

1 - le titre de la série,

2 - la description de la source et des fichiers où se trouve la série,

3 - un texte attaché comme commentaire à tout ou partie du titre de la série.

Ces informations figurent dans la base de données à des titres divers. En effet, en raison du mode d'exploitation d'une telle banque de données et du coût des stockages, traitements, l'information est répartie dans des fichiers qui s'inscrivent dans une hiérarchie fondée sur l'état général du système et la nature des supports caractérisés par leurs modes d'accès.

La base de données se trouve répartie sur deux niveaux : le niveau 1 est représenté par des archives et des descriptifs d'archives; le niveau 2 est constitué par l'information accessible en temps réel.

2.1.1. - les Archives

Les informations en archives sont réparties en deux catégories.

On considère les séries statistiques disponibles à l'extérieur du système, sur quelques supports que ce soit. Elles forment les archives externes. Les supports des archives externes sont divers et variés, depuis le document imprimé jusqu'aux supports lisibles par les ordinateurs. Une information archivée externe est décrite :

1) par le titre de la série statistique,

2) par une description de la source,

3) par une description de l'adresse. Celle-ci peut être un texte ou une description du fichier incluant la structure et l'adresse de la série. Cette adresse peut être physique ou constituée par la liste des clés et des valeurs des clés permettant l'accès à la série.

Les informations archivées internes sont, d'une part les descriptifs, d'autre part les séries statistiques, sous format standard, sur bandes magnétiques. Ces informations forment le niveau 1 de la base de données du SAD. Leur description est comparable à celle des informations externes.

Les archives internes sont obtenues par des programmes interface entre le SAD et les archives externes. Le rôle de ces interfaces est le changement de format, la vérification.

Le principe de constitution d'archives internes est celui de la non-redondance. Les informations sont stockées au niveau départemental ou régional, selon les cas, et avec le degré de finesse le plus élevé. Aucune information stockée n'est combinaison linéaire d'informations. Ce principe entraîne le stockage de très nombreuses séries qui ne sont souvent utilisées que pour la construction de séries réellement utilisées.

2.1.2. - L'information Active.

L'information ainsi désignée est constituée par une information extraite des archives internes, éventuellement agrégées à d'autres informations de même origine et disposée dans des fichiers à support d'accès

direct. Cette information est accessible en temps réel à l'utilisateur. L'information active entre dans le niveau 2 de la base de données

2.1.3. - L'information personnalisée.

Elle est constituée par des informations sélectionnées, ou introduites dans le système par l'utilisateur. Elles résident en mémoire centrale lors de l'utilisation et peuvent être conservées dans des fichiers à accès privatif stockées sur mémoire rapide.

L'information personnalisée n'entre pas dans la base de données, mais elle est gérée par le système.

La hiérarchie ainsi introduite a été élaborée en considérant :

1 - le point de vue de la rentabilité économique du système: On essaie de réduire le coût de stockage des données en ménageant au moment utile la possibilité pour l'utilisateur de travailler en temps réel.

2 - Le point de vue de l'utilisation de la banque de données: Devant le volume des données disponibles, l'utilisateur, en général ignorant cette richesse, procède en trois étapes :

- sélection de l'information utile,
- rassemblement et élaboration de celle-ci,
- traitement.

La hiérarchie introduite permet de construire le système offrant à l'utilisateur la souplesse désirée dans le déroulement de ces trois étapes et le laissant libre de ses décisions.

3 - le point de vue des interfaces externes. Le SAD en tant que banque de données économiques ne peut rester isolé des autres systèmes qui ont vu

le jour dans l'administration. En mettant en place des interfaces entre les archives internes du SAD et les fichiers externes, on peut ainsi les faire fonctionner de la façon la plus rentable sans alourdir le système mis à la disposition de l'utilisateur. La mise à jour de l'information active se fait toujours à partir des archives internes, qui, augmentant en volume, couvrent de plus en plus de besoins. En créant des archives internes structurées de façon simple et en les transformant en informations actives, on évite d'introduire dans les programmes de mises à jour de l'active la description de fichiers externes. Enfin, on atteint une plus grande modularité du système.

2.2. - La base de données.

La base de données est structurée en deux niveaux. Chaque niveau est constitué par les données et les informations qui en permettent l'accès, la recherche et le traitement.

Le niveau 1 est formé par un ensemble de fichiers étiquetés contenant, sous format SAD, les séries statistiques sur bandes magnétiques. Les articles sont repérés par un numéro.

Ce niveau est accessible par un programme d'interface entre les deux niveaux et par les programmes de gestion qui comportent :

- la création,
- la mise à jour,

soit par l'intermédiaire des interfaces externes, soit par l'intermédiaire de programmes spéciaux.

Le niveau 2 comprend :

1 - le fichier des informations actives, sous forme de séries statistiques, appelé fichier actif,

2 - les fichiers qui en assurent la description et l'accès,

3 - les fichiers des descriptifs des archives internes et externes, appelés fichiers documentaires,

4 - les fichiers de commentaires,

5 - des fichiers de services : tables, dictionnaires, etc ...

Tous ces fichiers sont stockés sur disques et sont consultés par le système d'utilisation.

La gestion de ce niveau est assurée par :

- l'interface entre les 2 niveaux,
- un programme de gestion des descriptifs des archives,
- un programme de gestion des fichiers commentaires,
- des programmes de gestion des fichiers de services.

2.3. - Fonctionnement Général du Système.

Les trois grandes fonctions assurées sont :

1 - création et gestion des archives internes,

2 - gestion du niveau 1, principalement constitué par l'interface niveau 1 - niveau 2,

3 - exploitation de la base de données au niveau 2.

2.3.1. - Création et Gestion des Archives Internes.

En ce qui concerne la création d'archives, SAD II considère cette

opération comme une opération assurée par des interfaces entre lui-même et d'autres systèmes.

Le premier a été l'interface SAD I - SAD II. Aux deux fichiers de SAD I, sont venus s'ajouter six autres fichiers de provenances diverses qui ont été mis sous les mêmes structures, portant le nombre de série-année à environ 70 000.

Le deuxième interface à créer est celui qui assure l'entrée dans la base de données des informations imprimées. Cette fonction correspond à celle assurée par le programme 750.A de SAD I. Le programme sera entièrement réécrit sous la forme d'un langage de définitions de tableaux. En effet, les données sont toujours imprimées sous forme de tableaux. Ces tableaux seront convertis sur un support lisible (cartes, bandes magnétiques) dans une structure décrite par des éléments d'un langage au niveau syntaxique et sémantique. En raison de la fréquence relative de certaines structures de tableaux, il est envisagé de procéder à la construction d'un macro-générateur permettant, dans la description des tableaux, de remplacer plusieurs instructions par une seule, construite par l'utilisateur.

D'autres programmes d'interface viendront s'ajouter aux précédents, selon les besoins, devenant permanents si une procédure d'échange d'informations est mise en place entre le SAD et un de ses fournisseurs d'informations.

Ces programmes ont pour but de construire des fichiers de séries statistiques. Il leur correspond des programmes de mises à jour des fichiers de descriptifs. La mise à jour de ceux-ci peut entraîner la restructuration des fichiers, afin de permettre un accès plus rapide en fonction du titre de la série.

La mise à jour des séries est surtout faite par ajout. Il y a peu de

modifications, et rarement des suppressions. Lorsque celles-ci se présentent, elles correspondent en fait à une modification globale de la série-année. Cette fonction de mise à jour par ajout est assurée par le programme de restructuration.

Ces programmes de gestion de la base de données au niveau archives sont destinés à fonctionner en "batch processing".

2.3.2. - Création et Gestion de la Base de Données Actives.

Le niveau 2 de la base de données comprend deux sortes de fichiers : les uns sont des fichiers de références aux informations en archive interne; les autres sont des fichiers doubles séries et descriptifs qui sont consultés et utilisés pour traitement.

La gestion des premiers est faite simultanément avec celle des fichiers archives internes. Il est envisagé pour la mise à jour de le faire selon les deux modes : "on line" pour ajouts peu volumineux, "batch" pour la création d'un sous-fichier de descriptifs correspondant à l'introduction en archive interne d'un fichier important de séries.

La création et la mise à jour des seconds est assurée par l'interface entre le niveau 1 et le niveau 2 de la base de données. Cet interface a pour rôle d'extraire des séries des archives internes et leurs descriptifs et de les fusionner au fichier actif. Cette opération est faite en temps différé, à partir des indications fournies par l'utilisateur. Elle peut conduire à des agrégations internes (régionalisation d'une série-année départementale), ou externes (sommation sur l'ensemble des caractères d'une caractéristique). Ces opérations sont faites par interprétation du titre des séries à amener au niveau 2.

La taille des fichiers actifs est variable dans le temps. Le contenu de ceux-ci est géré suivant un modèle qui prend en compte la date d'incorporation de la série-année dans ces fichiers et la fréquence d'utilisation.

En effet, on limite la taille des fichiers et il faut à chaque mise à jour envisager le remplacement de série-année peu utilisables à partir de ce jour par celles dont l'activation a été demandée.

Enfin, en ce qui concerne les fichiers de services, leur mise à jour peu fréquente après une période de démarrage du système, sera faite en "batch processing" dans un premier temps, puis "on line". Ceci concerne les dictionnaires, les fichiers d'informations décrivant les utilisateurs, la structure de la base de données. Le système de recherche de l'information est fondé sur l'utilisation du langage de description des données statistiques. La création du vocabulaire est faite en même temps que la création des descriptifs, le cas échéant. Aussi faut-il prévoir la mise à jour simultanée des fichiers concernés.

2.3.3. - Interrogation et Exploitation de la Base de Données.

Cette grande partie du système fonctionne sous 2 modes :

- mode "actif",

- mode "documentaire".

Le mode "actif" est conçu pour assurer les fonctions suivantes :

- sélection des série-année dans les fichiers actifs,
- accès aux série-année sélectionnées,
- traitements de ces séries,
- édition et sorties des résultats,
- gestion des fichiers personnalisés,
- recherche des commentaires liés aux séries sélectionnées.

La fonction de sélection des descriptifs commune aux deux modes porte sur les titres des séries. Elle peut dans le cas commentaire porter sur la source des informations et sur le fichier qui les contient. Ces deux interrogations ne sont pas les plus importantes. La recherche s'effectue en utilisant un langage très peu différent du français courant. Ce langage n'est pas délivré à l'utilisateur par le système, mais le système utilise le LDDS. L'interrogation est traduite dans ce dernier langage. Le système peut fournir à l'utilisateur les éléments du LDDS, à sa demande. L'interrogation, par titre, a alors deux aspects :

- l'aspect direct : la demande formulée par l'utilisateur est introduite en français. La traduction est effectuée. L'utilisateur peut demander à valider la question avant recherche.

- L'aspect "menu" : l'utilisateur demande que le système lui propose d'effectuer des choix. Ces choix s'effectuent au niveau des termes du langage. L'ensemble des termes choisis constitue une demande de titres syntaxiquement et sémantiquement correcte, rédigée en LDDS.

La recherche de commentaires liés aux descriptifs sélectionnés est assurée sur simple commande.

Les fonctions de recherche des séries, de traitement, d'édition, de gestion des fichiers personnalisés, sont assurées au moyen d'un langage conversationnel. L'utilisateur se trouve placé devant une zone de travail structurée de façon assez rigide, qu'il gère à sa convenance, en demandant des échanges avec la base de données commune ou personnalisée et les fichiers d'interfaces. Ces derniers sont créés par le système pour permettre l'utilisation de données issues, directement ou par traitement, des fichiers actifs. Ils sont utilisables par la suite en dehors du système par des programmes développés dans un contexte "on line" et interactif, ou fonctionnant en "batch" ou "remote batch".

2.3.5. - Programmes utilisateurs.

En dehors des possibilités étudiées ci-dessus, il est envisagé de donner des facilités d'accès et de recherche des séries aux programmes utilisateurs de nombreuses données en "batch ou remote batch processing". Ceci sera fait lorsque le système pourra disposer de mémoire de masse à très grande capacité pour le stockage des archives internes. Celle-ci peut toutefois être utilisée comme support de fichiers temporaires issus directement de fichiers en archive interne.

2.4. - Problèmes de Protection.

Les problèmes de protection se situent au niveau de l'accès aux données.

Une première limitation peut s'excercer au niveau des archives internes. Tout utilisateur se verra attribuer des autorisations d'accès aux archives au moyen de l'interface archive - fichiers actifs. Un problème pourtant se pose, en regard de la finesse des informations. En effet, on pourra interdire l'accès pour utilisation à un certain nombre d'informations, mais permettre la construction d'informations par agrégation, à partir d'informations interdites d'utilisation.

Cependant, ces cas sont encore peu nombreux dans l'usage prévu de SAD II.

En ce qui concerne les fichiers actifs, l'accès pourra être également modulé suivant le statut de l'utilisateur.

Enfin, l'accès aux fichiers personnalisés sera limité par un mot de passe, choisi par le créateur du fichier personnalisé.

En fait, le problème de l'accès à la base de données est surtout

subordonné à la destination finale du SAD et à l'usage qui en sera fait. Le problème est d'ailleurs lié à celui de la tarification. Un dictionnaire des utilisateurs sera créé et contiendra tous les renseignements utiles concernant chaque utilisateur en matière d'accès, tarifs, etc...

2.5. - Utilisation du Langage de Description des Données Statistiques.

Le langage de description des données statistiques est repris pour l'essentiel. La syntaxe reste pratiquement la même.

Un certain nombre de considérations d'ordre sémantique ont entraîné les modifications suivantes :

- toute caractéristique est liée à une partition,

- une caractéristique est dépendante d'un seul élément,

- une relation est définie par les éléments qu'elle relie et peut prendre plusieurs valeurs dans une liste.

Ceci entraîne que la suppression, dans un titre, d'une caractéristique, revient à engendrer les titres des séries correspondant à l'agrégation de toutes les séries dont les titres contiennent un des caractères de la caractéristique supprimée.

La relation entre caractéristique et élément est purement sémantique. Elle est introduite pour permettre une interprétation sémantique de l'interrogation.

Le codage du vocabulaire est réalisé de la façon suivante : tout terme est codé par un code alphanumérique à 7 caractères, et un entier sans signe, ou un symbole. Le symbole désigne une liste dans sa totalité ; l'entier sans signe pointe dans la liste. Le code alphanumérique pointe vers une liste de termes. Ce code est constitué d'un caractère alphabétique significatif de la

fonction syntaxique des termes de la liste, d'un entier sans signe à 3 chiffres qui établit la liaison entre élément et caractéristique ou élément, d'un deuxième entier sans signe à 3 chiffres qui repère la liste. Ces derniers entiers peuvent être nuls dans le cas où la liaison sémantique est indéterminée. Considérons l'élément "personnes", codé E.....(1) et deux caractéristiques attachées à cet élément. Elles sont codées : C OO3OOl et C Oll OOl. La totalité de la liste des caractères de la caractéristique est repérée par C OO3OOl (*) ; un caractère de celle-ci est repéré par C OO3OOl (2). Le code d'une caractéristique est de la forme : partie "numéro" partie "élément". est soit le caractère C, soit le caractère Q . La "partie numéro" repère la caractéristique par son numéro dans l'ensemble des caractéristiques liées à un élément dont le numéro est la partie "élément". On distingue deux catégories de caractéristiques : les caractéristiques qualitatives, et les caractéristiques quantitatives dont les caractères sont des intervalles. Elles sont distinguées dans la codage par C ou Q , comme premier caractère de leur code. Ainsi la caractéristique "par nombre de pièces" liée à l'élément "logement", codé E. ..... (5), est représentée par Q OOlOO5. Il existe des relations entre l'élément "personnes" et l'élément "logements". Considérons la première "habitant dans des". Celle-ci se code R OOlOO5 (1) Le code alphabétique significatif de la notion syntaxique de relation est R. Les deux parties entières sont les numéros d'éléments dans le sens de la relation. Une seconde relation pourrait être dans ce cas "propriétaire de" ; elle serait codée R OOlOO5 (2).

On peut alors, en utilisant un tel codage, tenter une traduction d'une phrase française en son équivalent en L.D.D.S. Le dictionnaire fait passer d'un mot français à une liste de code de termes. On construit alors toutes les combinaisons possibles en choisissant un code dans chaque liste associée à un mot français. Chaque combinaison est alors analysée du point de vue syntaxique. La présence de caractères, de caractéristiques et de relations permet de générer des éléments qui n'auraient pas été détectés par le dictionnaire. L'analyse sémantique conduit à sélectionner des combinaisons sémantiquement correctes, qui représentent des parties de titres.

Le codage du langage est en fait la représentation du thésaurus associé. Un autre effort, au point de vue syntaxique et sémantique, est fait au niveau du dictionnaire en y incluant également un thésaurus dans le codage interne des suffixes et stems dont la réunion représente un mot français.

La représentation de titre peut se faire par une structure arborescente dont l'origine est un terme d'élément et les extrémités les caractères des dernières caractéristiques. Un arc correspond à l'énoncé d'une caractéristique ou d'une relation. L'arborescence représente les possibilités d'agrégation par sommation, sur l'ensemble des arcs adjacents extérieurement à un sommet correspondant au niveau d'agrégation correspondant.

La liaison entre titres et séries se fait par pointeurs attachés aux extrémités des arbres.

Cette structure est adoptée pour la représentation des descriptifs des séries en archive interne. En effet, elle permet la sélection à tous les niveaux d'agrégation, et la génération de piles correspondantes dans le cas où l'utilisateur demande l'extraction de séries agrégées des archives internes. En outre, il est prévu de rechercher la pile traduisant les agrégations nécessaires qui optimise le temps de calcul.

En ce qui concerne la recherche de série-années dans les fichiers actifs, on utilise un fichier inversé sur les termes du titre. En effet, il est prévisible que l'économie de codage faite en adaptant une structure arborescente dépend de la nature des séries disponibles. Si toutes les séries correspondant à tous les caractères de toutes les caractéristiques existent dans les archives, on réalise une compression appréciable des descriptifs. Ce n'est certainement pas le cas dans les fichiers actifs, où les séries sont introduites à la demande et à des niveaux d'agrégations relativement élevés ; d'où il s'ensuit que leur descriptif comprend peu de termes et donc le fichier inversé peut ne pas être très volumineux.

SAD II est caractérisé par un langage d'interrogation différent du langage de définition des données. Celui-ci permet la structuration des fichiers de données et celle des fichiers de recherche, dont la structure est, ou bien arborescente, ou bien sous forme de liste.

2.6. - Caractérisation de SAD II.

La base de données est caractérisée par les deux niveaux. Les fichiers du niveau 1 sont sur bandes magnétiques, à structure simple,décrite par des fichiers à structure d'arbre appartenant au niveau 2 et sur disques. Ce deuxième niveau contient en outre des fichiers de données sur disque, et leurs descriptions. Les fichiers d'accès sont répartis en fichiers de textes (vocabulaires, commentaires), dictionnaires, à structure arborescente, en fichiers de services sous formes de tables.

La mise à jour des deux niveaux s'effectue en batch processing. Elle est décomposée en deux phases : constitution en archive interne de nouvelles données, mise à jour au niveau 2 des descriptions des données et restructuration au niveau 1, le cas échéant.

L'interrogation est faite "on line" sur les fichiers du niveau 2. Elle se fait par langage simple permettant une sélection des séries statistiques. Le traitement des séries est fait "on line" également au moyen d'un langage adapté aux objets manipulés (tableaux statistiques). Ce même langage permet la gestion de fichiers créés par l'utilisateur en vue de son application. Il donne la possibilité de procéder à des éditions et sorties diverses sur tout support en temps différé.

SAD II comprend des interfaces externes en entrée comme en sortie, et permet une utilisation de ses fichiers en "batch processing" pour des applications plus élaborées.

Le système SAD II sera vraisemblablement implanté sur CDC 6400.

Il requiert l'utilisation d'un "operating system" permettant l'utilisation en "batch" et ,"remote batch" ou "time-sharing".

Le système SAD II tel qu'il est actuellement conçu semble pouvoir rendre les services que des utilisateurs de toutes disciplines attendent d'une banque de données économiques localisées. Afin de mettre rapidement à leur disposition un système opérationnel, le système SAD II a été simulé par une réalisation plus légère, le système SAD II TS.

Chapitre III

LE SYSTEME SAD II T.S.

III - Le Système S A D II T.S.

Ce système a vu le jour comme une simulation d'une partie de SAD II limitée dans un premier temps au cas régional. Il fut conçu comme un système opérationnel ; l'observation qu'on pouvait faire de son utilisation par des utilisateurs non informaticiens et travaillant en "vraie grandeur" devait donner la possibilité d'infléchir, le cas échéant, les orientations prises dans SAD II. Enfin, il devait permettre de remplacer avantageusement SAD I au service de la Délégation à l'Aménagement du Territoire et à l'Action Régionale, d'autant plus que celle-ci se voyait privée de l'utilisation du programme principal de SAD I (750-CORSE) par suite de modification de configuration de l'ordinateur utilisé.

Ce système a été implanté en partie sur l'ordinateur de SAD I (CII 90-80) pour ce qui concerne la gestion des données, en partie sur XDS 940 pour ce qui concerne la "Recherche de données"et les " Traitements". En effet, on considère les fichiers de SAD I comme les archives internes de SAD II TS. Le fichier actif de SAD II TS est implanté sur les disques à grande capacité du système 940 XDS. L'interface entre les archives externes et internes est constituée par l'ensemble des programmes d'enregistrement et de mise à jour de SAD I. L'interface entre le fichier actif et les archives externes est constituée par deux programmes, l'un sur C I I 90.80, qui assure le transfert de l'information d'une configuration à l'autre, le second sur XDS 940, qui lit les fichiers intermédiaires et les transforme en des fichiers exploitables par le système SAD II TS. Cette organisation permet un fonctionnement correct de la banque de données tant que les demandes d'accès aux informations en archives externes et internes ne sont pas trop fréquentes ni trop urgentes. SAD II TS recevra encore bien des améliorations mais il n'est pas envisagé de le compléter par la partie gestion des données qui lui manque. En effet, il est préférable de transférer cette fonction à SAD II, dont SAD II TS apparaitrait alors comme un sous-système disponible sur un ordinateur (XDS 940) d'accès peu coûteux.

Le système permet alors d'être implanté sur plusieurs ordinateurs XDS 940 disponibles en France.

Le système SAD II TS se divise en deux sous-systèmes gérés par un module unique accessible de façon extrêmement simple à l'utilisateur. Celui-ci se voit offrir des loyens de traitements des descriptifs des séries actives, d'une part, et des moyens de traitements des séries actives elles-mêmes d'autre part. Ces traitements, effectués en time-sharing, sont obtenus par un dialogue entre le système et l'utilisateur. On atteint ainsi l'un des objectifs du SAD, le mode conversationnel et le traitement à distance. Le système, sans être en bibliothèque générale, est accessible à tout abonné du système des prestations de service.

Le principe de fonctionnement est le suivant : l'utilisateur dispose en mémoire centrale d'une zone de travail structurée qu'il atteint par des références simples. Il dispose d'un langage qui lui permet les échanges d'informations, le traitement, les éditions

Les échanges d'informations sont doubles. Ils peuvent se produire depuis les fichiers personnalisés de l'utilisateur vers la zone de travail et vice-versa, ainsi que des fichiers fondamentaux vers la zone de travail. Le système permet la création par l'utilisateur de fichiers personnalisés qui s'inscrivent au "File Directory" de son numéro de compte et ne font pas, à proprement parler, partie du système SAD II TS. Celui-ci en permet la création et la gestion. Le traitement des informations a lieu dans les deux sous-systèmes, en mémoire centrale dans la zone de travail. En ce qui concerne les traitements des données statistiques, ceux-ci sont d'ordre arithmétiques et statistiques. L'utilisateur a la charge de gérer la zone de travail. Il peut procéder à des éditions sous formes diverses sur le terminal, ou dans des fichiers symboliques qui, "sauvés sur disques, peuvent ensuite être édités sur carte ou bande magnétiques, par un travail en temps différé et effectué "in situ". En ce qui concerne les traitements des descriptifs de séries, ceux-ci sont d'ordre logique. L'utilisateur a également la charge de gérer

la zone de travail des références ou séries. Il peut se livrer à des travaux d'éditions de descriptifs de séries sur le terminal ou dans des fichiers symboliques sur disques en vue d'édition sur d'autres supports.

1. - Utilisation des ressources du système XDS 940 par SAD II TS.

Le système Time-Sharing XDS 940 se présente comme un ensemble de sous-systèmes reliés à un noyau central, l'exécutif. Celui-ci a pour tâches principales, vis-à-vis de l'utilisateur, l'appel des sous-systèmes, la gestion du dictionnaire des fichiers, l'exécution des programmes exécutables, la gestion des interruptions qui font sortir des sous-systèmes. Ceux-ci sont au nombre de 5 : le BASIC, le FORTRAN IV conversationnel, le QUED (éditeur de texte), le CAL et enfin le FØRTRAN qui a été utilisé pour SAD II. Celui-ci n'est pas conversationnel, mais permet de créer et gérer des fichiers binaires. Dans ce sous-système divisé en deux : compilation d'une part, chargement et exécution d'autre part, l'utilisateur dispose d'une mémoire d'environ 12 K mots de mémoire centrale. Les mots sont à 24 bits. Ils permettent de stocker 3 caractères à 8 bits. Le sous-système permet la création et l'utilisation de programmes chaînés. Le chaînage n'est pas limité à un seul élément. Les fichiers contenant les chaînages sont créés par le module chargement et exécution. Il en est de même pour les programmes exécutables. Ceux-ci sont exécutés à la suite d'une commande "GØTO" de l'exécutif qui transmet le contrôle au programme exécutable contenu dans le fichier dont le nom figure dans la commande. Enfin, les fichiers d'un utilisateur sont accessibles par un autre utilisateur pourvu qu'ils aient un nom comportant le caractère @ . De la sorte, il est possible à un utilisateur d'accéder à SAD II TS par une commande GØTO portant sur le programme unique initialisateur. (cf. (3)).

1.2. - Structure du système : les Fichiers.

La base commune est constituée par :

1) les fichiers de données actives.

Ils contiennent l'information statistique sous forme de série-année régionale.

Les fichiers sont binaires et exploités de façon séquentielle. La structure de l'enregistrement est la suivante :

1er mot	:	numéro de série-année
2ème mot	:	facteur d'échelle (cf. 5.1.2.)
3ème mot	:	type
4ème mot	:	vide ou valeur France entière
du 5ème mot au 26ème mot	:	les 22 valeurs correspondant aux 22 régions
27ème mot	:	numéro de série-année de l'article suivant (cas du dernier article : mot nul).

Les articles sont donc de longueur fixe. Le 2ème mot contient exactement le facteur d'échelle x défini en 5.1.2. Le type est une information qui précise la sommabilité ou la non-sommabilité de la série-année et la validité de l'information pour la Corse et la Provence-Côte d'Azur. Le type proprement dit vaut 1 si l'information pour la Corse est incluse dans l'information pour la Provence-Côte d'Azur, 2 si elle manque complètement, 3 si elle est disponible indépendamment de l'information pour la Provence-Côte d'Azur, donc dans le 26ème mot. L'information type de l'enregistrement est égale au type si la série est sommable, au type plus 10 si la série est non sommable.

2) un fichier table dont la structure est la suivante :
L'enregistrement est constitué par 4 mots binaires.

1er mot	:	numéro de série dans le fichier archive (SAD I)
2ème mot	:	numéro de série dans le fichier actif (SAD II TS)

3ème mot : indicateur de validité

4ème mot : adresse du descriptif.
Cette adresse est constituée par un numéro de fichier dans une liste et l'adresse dans le fichier qui est inférieure ou égale à 999

3) Les fichiers des descriptifs

Ils contiennent les descriptifs sous forme binaire. Ils ont la structure suivante par article de longueur variable :

ler mot : longueur de l'enregistrement

2ème mot : le numéro de série active

3ème mot : index de la série dans le fichier catalogue

4ème mot : libellé de 12 caractères donnant les années pour
7ème mot : lesquelles la série figure dans les fichiers.

du 8ème mot libellé du descriptif de la série. Le nombre de
au même mot : mots est limité à 24 ($n \leqslant 30$).

Le premier mot du fichier représente la longueur en mot du fichier. L'index de la série dans le fichier catalogue est le numéro du chapitre suivi du numéro de la section, suivi du numéro de paragraphe auxquels appartient la série.

4) le fichier catalogue

Ce fichier est un fichier en arbre dont les noeuds représentent des niveaux dans un catalogue des données. Il est structuré en chapitres, sections, paragraphes. Des noeuds pendants renvoient à une adresse dans les fichiers des descriptifs. L'ensemble des descriptifs d'un même chapitre, d'une même section, d'un même paragraphe,

sont enregistrés dans les fichiers précédents à partir de cette adresse.

La structure d'un article du fichier catalogue est la suivante :

1er mot	:	longueur de l'article,
2ème mot	:	référence du noeud à un niveau donné,
3ème mot	:	adresse du premier article du niveau inférieur,
4ème mot au nième	:	libellé limité à 60 caractères du descriptif du noeud.

5) Divers petits fichiers annexes qui constituent des tables de noms de fichiers, de références,

6) un fichier dictionnaire.

Ce dictionnaire assure le passage d'une chaîne de caractères à une référence (pointeur vers le fichier inversé);l es chaînes de lettres formant soit des stems, soit des suffixes. La structure du fichier est arborescente. A chaque caractère d'une chaîne stockée dans un mot correspond un mot où est stocké un pointeur qui repère un caractère alternant. La fin de la chaîne est repérée par un caractère spécial. Il lui est associé un pointeur. Le pointeur de stem est une référence dans le fichier inversé avec une partie "syntaxique" et une partie "sémantique". Le pointeur de suffixe a aussi deux parties. La concordance des parties syntaxiques des pointeurs de stem et de suffixe permet d'assurer que la réunion stem suffixe est un mot français.

Cette partie est décrite dans (2).

7) Un fichier inversé sur le vocabulaire des descriptifs de série.

Ce fichier a la structure d'article suivante :

1er mot : référence d'un mot du dictionnaire.

2ème mot : nombre de numéros de série

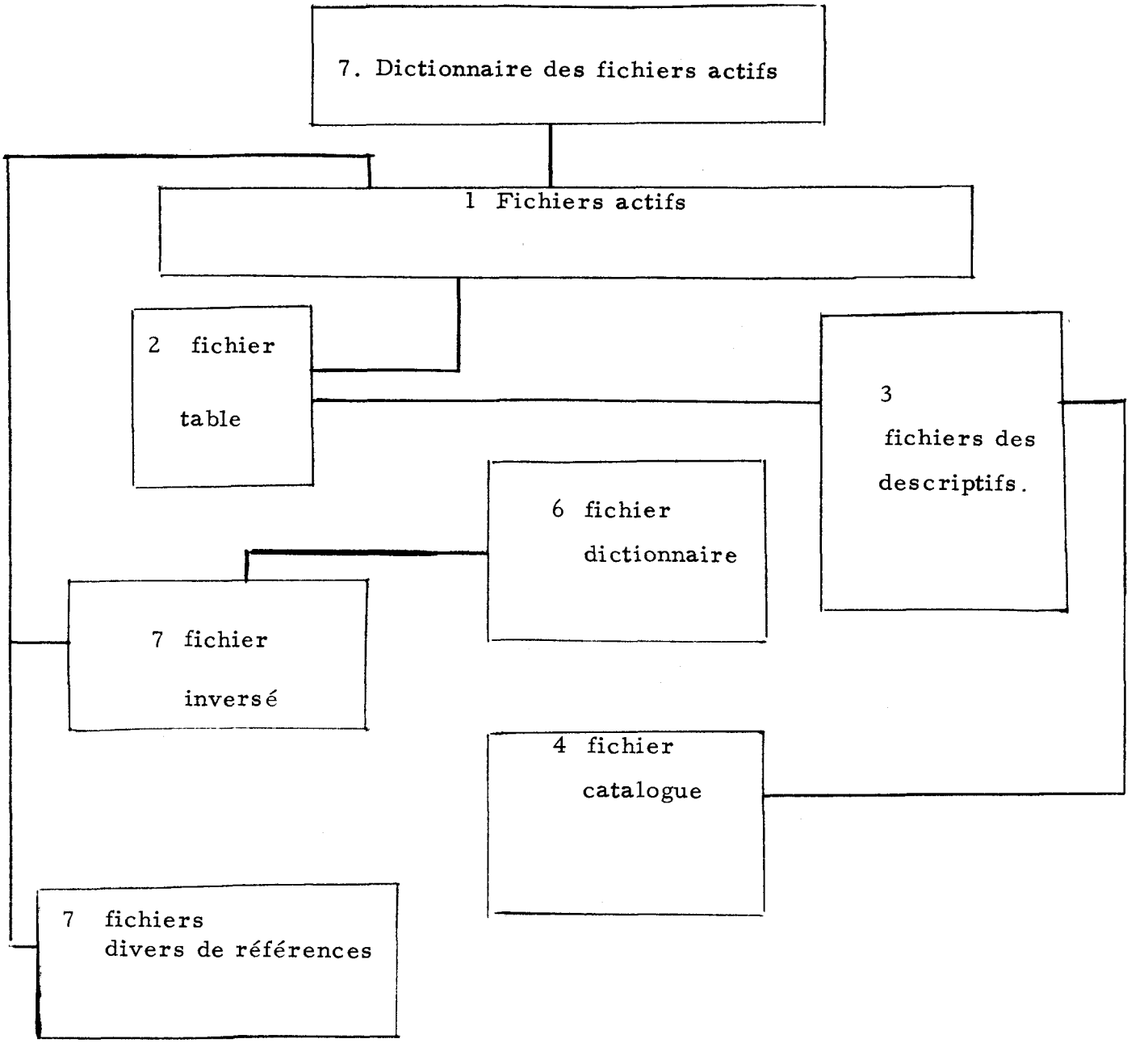
du 3ème mot
au nème mot : liste des numéros de séries

Les structures de fichiers sont prévues pour l'utilisation de l'accès direct. Celui-ci n'est pas encore mis en service. On a, pour le simuler, construit un sous-programme de lecture séquentielle du fichier dans un buffer de 255 mots, taille d'un data block. Ces data blocks constituent la structure des fichiers du système XDS 940. On optimise ainsi le compromis espace mémoire-temps d'accès à l'information. Cependant, les fichiers de données sont encore explorés séquentiellement. On prévoit de remplacer cette méthode d'accès par une méthode genre séquentiel indexé. L'adresse de la première série-année serait d'une série fournie par le fichier "table".

Les fichiers sont consultés par les deux sous-systèmes.

La recherche d'une série-année et de son titre pour leur transfert en mémoire centrale se fait en connaissant une référence à cette série-année. Cette référence est constituée par un entier à 6 chiffres dont le premier est le numéro du fichier actif dans la liste des fichiers actifs, liste qui est dans un fichier symbolique. Les trois caractères suivants forment le numéro de série, les deux derniers la référence d'année. L'utilisation d'un fichier liste de références de fichiers s'est imposée pour permettre des limitations d'accès à de larges ensembles de données, et toutes sortes d'opérations sur les références plutôt que sur les fichiers. Enfin elle permettait de ne pas inclure dans la programmation le nom des fichiers, donc de pouvoir les changer. La recherche du titre de la série à partir du numéro de série se fait par consultation de la table et lecture dans l'un des fichiers de descriptifs.

Le dictionnaire est utilisé pour la transformation d'une phrase écrite en français, en une liste de mot-clés repérés par leur numéro dans le dictionnaire. Ce numéro sert d'entrée dans le fichier inversé qui fournit des numéros de séries dont le titre contient le mot-clé. Le numéro de série renvoie au titre de la série.



Structure de la base commune de données

Structure du système : les programmes d'exploitation de la base de données.

Le programme P initialisateur propose à l'utilisateur le choix entre le sous-système "Recherche de Titres" ou le sous-système "Traitement des séries" : il propose également la sortie sur le terminal de la notice d'emploi du système. Le choix de l'utilisateur entraîne le chaînage sur le programme principal du sous-système choisi. Dans chacun des sous-système, l'utilisateur a, à sa disposition, une zone de travail. Cette zone de travail est conservée lors d'un changement du sous-système demandé par l'utilisateur. Elle est restaurée à un nouvel appel du sous-système. L'utilisateur introduit ses demandes de traitements grâce à un langage adapté à chacun des sous-systèmes. Chaque instruction est exécutée après décodage, et le programme est mis en position d'attente d'une nouvelle instruction.

Les chaînages à partir du programme P se font soit vers le programme L1, programme principal du programme sous-système "TRAITEMENTS", soit vers le programme S1, programme principal du système "RECHERCHE", soit vers N programme d'édition de la notice mode d'emploi.

Le chaînage principal d'un sous-système assure les fonctions :

1) écriture de la zone de travail sur disque lorsque l'utilisateur a demandé l'autre sous-système,
2) restauration de la zone de travail avant l'entrée dans le sous-système,
3) décodage partiel ou total de l'instruction,
4) exécution de certaines instructions,
5) chaînage vers les programmes secondaires.

Un programme secondaire assure les fonctions :

1) décodage partiel ou total des instructions,
2) exécution de certaines instructions,
3) chaînage vers le programme principal ou le programme secondaire.

La partie initialisation des programmes principaux est utilisée pour effectuer des tests permettant de connaître l'état de la mémoire et la position du programme

vis-à-vis de l'utilisateur.

La répartition des primitives d'exécution a été faite entre les programmes principaux et secondaires de façon à réduire le nombre de chaînages donc le temps de réponse.

4. - Le langage d'utilisation

Le langage comprend un langage de contrôle du déroulement du programme, un langage de traitement des séries statistiques, un langage de recherche et de traitement des titres de séries. Les deux langages liés aux sous-systèmes sont spécifiques de ceux-ci, tandis que le langage de contrôle leur est commun. Les instructions sont frappées après demande de la part du programme imprimée sur le terminal.

4.1. - Le langage de contrôle

Les instructions ont la structure suivante :

identificateur d'opération

l'identificateur d'opération est une chaîne à 3 caractères.

Les instructions sont les suivantes :

l'Instruction :

FIN	Termine une séance de travail.
PAU	Suspend l'exécution du dialogue .
DAT	Imprime la date et l'heure sur le listing du terminal.
TEM	Donne le temps écoulé en minutes depuis le début de la séance de travail.
SYS	Transfert le contrôle à l'autre sous-système.
NOT	Permet d'obtenir des informations sur le mode d'emploi et rend le contrôle au sous-système en cours.

4.2. - Le langage de recherche de titres.

L'utilisateur dispose d'une zone de travail de 10 colonnes et 300 lignes. Chaque colonne peut contenir 300 références de séries. Celles-ci sont obtenues, soit par une recherche, soit par opérations logiques entre colonnes Un pointeur de colonne est géré automatiquement par le système et pointe la colonne "courante" Chaque colonne est dotée d'un descriptif permettant à l'utilisateur de mémoriser un commentaire sur le contenu de la colonne.

La description des instructions se fera avec la convention suivante :

un numéro de colonne sera représenté par X, Y, Z ; s'il est placé entre crochets, cela indique qu'il est facultatif.

1) Instruction PHR :

Elle s'écrit : PHR [X]

Cette instruction permet à l'utilisateur d'introduire une description en français des titres de séries dont il désire voir les références apparaître dans la colonne X, ou si X est omis dans la colonne qui suit la colonne "courante". Il aura alors à fournir à l'issue de la recherche un descriptif de colonne. Le pointeur est augmenté de 1 si on a utilisé la colonne "courante".

2) Instruction EDI.

Elle s'écrit : EDI [X]

Elle a pour effet de lister les références contenues dans la colonne X ou dans la colonne "courante".

3) Instruction TIT.

Elle s'écrit : TIT [X]

Elle entraîne l'impression des titres de séries dont les références sont contenues dans la colonne X ou dans la colonne "courante".

4) Instruction LIR.

Elle s'écrit LIR [X , Y]

Cette instruction entraîne l'impression, sur la liste du terminal, du descriptif de la colonne X ou de la colonne "courante", ainsi que le nombre de références contenues.

5) Instruction OU

Elle s'écrit OUb X , Y, [Z]. (1)

Cette instruction réalise la réunion logique des références des colonnes X et Y. Cet ensemble de références est placé dans la colonne Z ou dans la colonne qui suit la colonne "courante". Dans ce dernier cas, le pointeur est augmenté d'une unité. L'utilisateur fournit un descriptif.

6) Instruction ET

Elle s'écrit ETb X, Y, [Z] (1)

Cette instruction réalise l'intersection des ensembles de références contenues respectivement dans les colonnes X, Y. L'ensemble résultant est contenu, soit dans la colonne Z, soit dans la colonne qui suit la colonne "courante". Dans ce dernier cas, le pointeur est augmenté d'une unité. L'utilisateur fournit alors un descriptif.

(1) b signifie le caractère "blanc".

7) Instruction SAU

Elle s'écrit SAU X, Y, [Z]

Cette instruction a pour effet de constituer un ensemble de références contenues dans la colonne X et non contenues dans la colonne Y. L'ensemble ainsi construit est contenu soit dans la colonne Z, soit dans la colonne qui suit la colonne "courante" et le pointeur dans ce dernier cas est augmenté d'une unité. L'utilisateur fournit un descriptif.

8) Instruction TRA

Elle s'écrit TRA X, [Y]

Le contenu de la colonne X est alors transféré dans la colonne Y ou dans la colonne qui suit la colonne "courante". Dans ce dernier cas le pointeur est augmenté d'une unité.

9) Instruction SUP

Elle s'écrit SUP.

Cette instruction a pour effet de diminuer le pointeur de colonne "courante" d'une unité.

10) Instruction NCL

Elle s'écrit NCL.

Cette instruction provoque sur la liste du terminal l'impression de la valeur du pointeur.

11) Instruction RAZ.

Elle s'écrit RAZ [X]

Cette instruction ramène le pointeur à la valeur X ou à O si X est omis.

L'ensemble de ces instructions permet à un utilisateur de faire une recherche de titres et de se composer des ensembles de références à l'aide d'opérations simples. Ces opérations permettent d'affirmer une recherche, de rassembler des références utiles provenant de recherches menées sur des concepts différents.

4.3. - Le langage de traitement des séries numériques.

La zone de travail est structurée comme un tableau de 23 lignes et 4O colonnes. Cette zone permet de traiter simultanément 4O série-année régionales. Chaque colonne est pourvue d'un titre, et d'une variable de "type" (cf. 5.1.). L'utilisateur dispose d'instructions d'échanges, d'instructions arithmétiques, d'instructions de gestion de la zone de travail, d'instructions de sortie et d'édition des informations de la zone de travail.

La description des instructions est faite avec les conventions d'écriture suivantes :

a) une lettre minulcule représentera un entier signé ou non

b) un numéro de colonne est représenté par une lettre majuscule.

c) tout caractère entre crochets indique que l'élément référencé par ce caractère est d'un emploi facultatif dans l'écriture d'une instruction.

1 - les instructions d'échange.

Les échanges d'information ont lieu :

- des fichiers actifs vers la zone de travail,

- de la zone de travail vers le fichier personnalisé,

- d'un fichier personnalisé, vers la zone de travail.

Ces instructions permettent à l'utilisateur d'accéder, d'une part à la base commune de données, d'autre part à sa propre base de données et de créer, gérer les fichiers qui composent cette dernière base de données. Celle-ci n'est connue du système que par l'intervention de l'utilisateur qui en a la responsabilité à ses frais.

1.1. - Introduction PER (écriture dans un fichier personnalisé).

Cette instruction s'écrit : PER

Elle a pour effet de provoquer l'écriture des 20 premières colonnes de la zone de travail dans un fichier personnalisé. Le nom du fichier est demandé à l'utilisateur. Celui-ci a la possibilité de donner un titre à son fichier. Ce titre a une longueur maximum de 72 caractères. Il est introduit par l'utilisateur en réponse à la question TITRE du FICHIER ? imprimée sur la liste du terminal.

1.2. - Instruction CHP (lecture d'un fichier personnalisé - chargement d'un fichier personnalisé)

Cette instruction s'écrit CHP [X]

Cette instruction permet le chargement d'un fichier personnalisé de 20 colonnes à partir de la colonne X ou de la colonne 1, si X est omis. Le nom du fichier est alors demandé et introduit par l'utilisateur. Le chargement se termine par l'impression du titre du fichier.

1.3. - Instruction CHA (chargement à partir des fichiers Actifs)

Cette instruction peut s'écrire sous 3 formes différentes.

Première forme :

CHA x, X

x est un numéro de série à 4 chiffres. X un numéro de colonne. Cette instruction provoquera le chargement de toutes les série-année de la série référencée par x dans la zone de travail. Le titre de chaque série-année sera également chargé avec la référence à l'année entre parenthèses. Le programme procèdera à la fin du chargement, à l'impression de l'implantation en zone de travail des série-année chargées. Celles-ci sont chargées à partir de la colonne X.

Deuxième forme :

CHA x, y, X.

x est un numéro de série à 4 chiffres, y une référence d'année à 2 chiffres. X un numéro de colonne. L'instruction a pour effet le chargement de la série-année x y dans la colonne X. Le titre de la série-année est également chargé. Le programme termine le chargement par l'impression de l'implantation en zone de travail.

Troisième forme :

CHA x, y, [p,] q , X

x et y sont des numéros de série, p est un entier positif appelé pas qui est pris par convention égal à 1 s'il est omis, q est une référence d'année à 2 chiffres, X un numéro de colonne. L'instruction provoque le chargement à partir de la colonne X, des série-année correspondant à l'année q et appartenant aux séries dont les numéros appartiennent à la liste définie par la borne inférieure x, la borne supérieure y et le pas de progression p. Le chargement s'effectue à partir de la colonne X dans des colonnes consécutives dans tous les cas.

Les numéros de série-année x et y doivent s'écrire avec le même premier chiffre. Le nombre de colonnes chargées par une seule instruction CHA ne peut

excéder 30.

L'instruction CHA permet, grâce à ses trois formes, d'effectuer des chargements de données, à partir des fichiers actifs, qui résolvent facilement les problèmes courants de l'utilisateur.

2 - Les instructions de traitements arithmétiques.

Ces instructions permettent des calculs arithmétiques entre colonnes. L'utilisateur peut provoquer l'impression des résultats sur la liste du terminal. Cette impression est sommaire et simplement destinée à permettre à l'utilisateur de prendre connaissance du résultat dans quelque colonne qu'il soit rangé, sans pour cela demander l'impression de cette colonne par une instruction de sortie. Pour cela, l'utilisateur complètera le code alphabétique d'instruction par le caractère /.

2.1. - Instruction ADD (Addition de colonne).

Elle s'écrit : ADD X , Y, [Z]

Cette instruction provoque l'addition des contenus des deux colonnes X et Y et range le résultat dans la colonne Z ou la colonne X si Z est omis. Un nouveau titre de colonne est alors demandé par le programme.

2.2. - Instruction SOU (soustraction de colonnes)

Elle s'écrit : SOU X, Y , [Z]

L'instruction provoque la soustraction du contenu de la colonne Y du contenu de la colonne X. Le résultat est rangé dans la colonne Z ou la colonne X si Z est omis. Un nouveau titre de colonne est alors demandé par le programme.

2.3. - Instruction PRO (Produit de colonnes)

Elle s'écrit : PRO X, Y [, Z]

L'instruction entraîne le calcul ligne à ligne du produit des contenus des colonnes X et Y. La colonne résultat est Z ou X si Z a été omis. Un nouveau titre de colonne est alors damandé par le programme.

2.4. - Instruction QUO (quotient de colonnes).

Elle s'écrit : QUO X, Y, [, Z]

L'instruction entraîne le calcul, ligne à ligne, du quotient des contenus des colonnes X et Y. Le résultat est rangé dans la colonne Z ou la colonne X si Z a été omis. Un nouveau titre de colonne est alors demandé par le programme.

2.5. - Instruction RAT (ratio de la colonne)

Elle s'écrit : RAT X, Y [, Z]

Cette instruction procède comme l'instruction QUO. Le résultat est celui de l'instruction QUO multiplié par 100. Un nouveau titre de colonne est demandé par le programme.

2.6. - Instruction TAU (variation relative en pourcentage entre 2 colonnes).

Elle s'écrit TAU X, Y [, Z]

Cette instruction effectue la différence entre les contenus des colonnes X et Y et divise celle-ci par le contenu de la colonne X, multiplie le résultat par 100.

Le résultat final est rangé soit en Z, soit en X si Z a été omis. Un nouveau titre de colonne est demandé par le programme.

2.7. - Instruction MUL X [, Y]

Cette instruction provoque la multiplication de tous les éléments de la colonne X par un scalaire. Celui-ci est demandé à l'utilisateur par le programme. Le résultat est rangé dans la colonne Y ou la colonne X si la référence Y a été omise.

Un nouveau titre de colonne est alors demandé par le programme.

2.8. - Instruction PCT (pourcentage par rapport à la dernière ligne d'une colonne).

Elle s'écrit : PCT X [, Y]

Cette instruction calcule les éléments de la colonne X en pourcentage de la valeur correspondante au niveau "France", c'est-à-dire à la valeur de la dernière ligne de la colonne X. Le résultat est rangé dans la colonne Y ou dans la colonne X si la référence Y a été omise. Un nouveau titre de colonne est alors demandé par le programme.

3 - Les Instructions Statistiques.

3.1. - instruction STA (calcul des principaux éléments statistiques d'une colonne).

Elle s'écrit : STA X [, Y, ..., ... Z]

Cette instruction entraîne le calcul pour les colonnes X, Y..., Z des éléments suivants :

- Valeur maximum de la colonne,
- valeur minimum " ,
- moyenne des valeurs " ,
- écart-type.

Les éléments sont imprimés sur la liste du terminal avec pour les deux premiers le libellé de l'unité territoriale à laquelle la valeur est attachée.

3.2. - Instruction COR (corrélation entre 2 colonnes).

Elle s'écrit : COR X, Y.

Cette instruction provoque le calcul du coefficient de corrélation entre les colonnes X et Y et imprime le résultat sur la liste du terminal.

3.3. - Instruction PRJ (Projection dans le temps à partir de certaines colonnes).

Elle s'écrit : PRJ X, Y, Z , U,.....W.

Cette instruction permet d'associer aux colonnes X, Y.....W des dates et de calculer pour chaque unité territoriale, par ajustement linéaire ou exponentiel, la valeur correspondant à une date donnée.

Le programme demande pour chaque colonne la valeur correspondante de la date (année) ; puis il demande la date pour laquelle on effectue la projection ainsi que la colonne où sera stocké le résultat ainsi que son titre. L'utilisateur devra aussi choisir la méthode (linéaire ou exponentielle) en réponse à une question du programme.

4 - Instructions de gestion de la zone de travail.

4.1. - Instruction TRA (Transfert du contenu d'une ou plusieurs colonnes).

Cette instruction peut prendre 2 formes.

lère forme :

TRA X , Y

Le contenu de la colonne X est transféré dans la colonne Y. Le titre de la colonne Y est alors celui de la colonne X.

2ème forme :

TRA X, Y, [p,] Ƶ [, q]

Le contenu des colonnes dont les numéros appartiennent à la liste définie par la borne inférieure X et la borne supérieure Y avec un pas p (ou 1, si p est omis), est transféré dans les colonnes dont les numéros appartiennent à la liste définie par sa borne inférieure Ƶ avec un pas q (ou 1, si q est omis). Le transfert des titres de colonnes est également assuré.

4.2. - Instruction MOD (modification d'une valeur dans une colonne).

Elle s'écrit MOD X [, Y]

Cette instruction provoque de la part du programme une demande double :

- Demande du numéro de la valeur à modifier,
- nouvelle valeur.

Le programme procède à l'impression de l'ancienne valeur.

4.3. - Instruction COL (Introduction de 23 valeurs dans une colonne).

Elle s'écrit COL X [Y, ... Z]

Cette instruction permet d'introduire une série de 23 valeurs pour chacune des colonnes X, Y . . . . . Z, soit en frappant ces valeurs, soit par l'intermédiaire d'un ruban perforé.

4.4. - Instruction ECR (Introduction d'un nouveau titre de colonne).

Elle s'écrit ECR X[, Y, ... , ... , Z]

Cette instruction permet à l'utilisateur d'introduire des nouveaux titres pour les colonnes X, Y, . . . Z.

5 - Les Instructions d'édition et de sortie.

Ces instructions ont pour but de permettre l'édition de tout ou partie des informations de la zone de travail sous des formats divers et sur des supports divers. Ces éditions peuvent prendre plusieurs aspects suivant les désirs de l'utilisateur.

Les sorties permettent l'impression sur le terminal de résultats limités donnant ainsi les moyens à l'utilisateur de contrôler le déroulement des instructions et de décider de la poursuite des travaux.

Les instructions d'édition.

Les éditions peuvent se faire sur la liste du terminal, sur imprimante ou sur bande magnétique ou sur carte. Dans ces derniers cas, l'édition se fait d'abord sur un fichier symbolique du système 940. La conversion des supports aura lieu alors, in situ, sur demande de l'utilisateur à l'opérateur du système. Pour obtenir une édition sur fichier symbolique, il suffit d'ajouter un code alphabétique d'instructions.

le caractère ✕. Auparavant, l'utilisateur aura à créer certains fichiers (cf. Guide du terminal).

5.1. Instruction IMP.

Elle s'écrit IMP X [, Y, ... Z]

Cette instruction provoque l'édition d'un tableau dont les colonnes auront pour contenu celui des colonnes X, Y, ... Z de la zone de travail. Pour cela le programme demandera à l'utilisateur des éléments de mise en page. La première question posée concerne le choix de l'utilisateur entre une impression de toutes les régions ou seulement l'impression de quelques régions dont il fournira les références.

Le programme demandera le nombre de chiffres après la virgule (Format) pour chacune des colonnes du tableau. Puis il demandera des éléments pour composer des en-tête de colonnes. Ces en-tête peuvent porter sur les n colonnes du tableau ou moins ; leur longueur ne peut excéder n ✕ 11 - 3 caractères pour n colonnes. Leur position est définie dans une ligne par le numéro de la première colonne. Une ligne est terminée quand la dernière colonne a été dotée d'un en-tête. L'utilisateur a droit à 5 lignes d'en-tête. Pour arrêter la définition des en-tête, il frappe la touche "return" à la question titre.

Le programme imprime à la demande un commentaire du tableau constitué par les titres des colonnes de la zone de travail qui sont éditées dans le tableau.

5.2. - Instruction ORD (Edition d'un tableau dont la première colonne est ordonnée par valeur décroissante).

Cette instruction s'écrit : ORD X [, Y, ... Z]

Cette instruction provoque l'édition d'un tableau comme l'instruction IMP. Les lignes du tableau seront alors éditées dans l'ordre décroissant des valeurs de la colonne X.

Le dialogue de définition du tableau est identique à celui provoqué par l'instruction IMP.

5.2. - Instruction DIA (Edition d'un diagramme).

Elle s'écrit : DIA X , Y.

Cette instruction provoque l'édition d'un diagramme en bâton construit sur les valeurs de la colonne X et éventuellement de la colonne Y. La longueur des segments est proportionnelle aux valeurs contenues dans la colonne X (ou Y).

5.3. - Instruction CAR (Edition d'une carte).

Elle s'écrit : CAR X.

Cette instruction entraîne l'édition d'une carte de France divisée en régions où s'inscrivent les valeurs correspondantes de la colonne X.

5.4. - Instruction EXT (Edition de données dans un fichier symbolique).

Elle s'écrit EXT X,[Y, ... , Z]

Cette instruction provoque l'édition du contenu de la colonne X (et des colonnes Y, ... Z, le cas échéant), dans un fichier symbolique et sous un format standard, sans aucun libellé. Ce fichier est destiné à jouer un rôle d'interface entre le système et les programmes des utilisateurs.

Instruction de Sortie.

5.6. - Instruction CAT.

Elle s'écrit : CAT [p, q, r]

Cette instruction permet d'établir un dialogue pour l'exploration du fichier catalogue. p est un numéro de chapitre, q un numéro de section du chapitre choisi, r un numéro de paragraphe de la section,choisi. les choix s'effectuent

au niveau de l'instruction, soit en réponse à une question après impression des libellés correspondants. A la fin de l'exploration, le programme procède à l'impression de la liste des titres de séries avec leurs références et les années pour lesquelles elles figurent dans le fichier actif.

5.6. - Instruction LIR (Sortie de titres de colonnes).

Elle s'écrit : LIR X [, Y, ...Z]

Cette instruction entraîne l'impression sur la liste du terminal des titres de la colonne X (et des colonnes Y, ...Z.

5.7. - Instruction COM (sortie de 20 titres de colonnes).

Elle s'écrit : COM X

L'instruction a pour effet de provoquer l'impression des titres des 20 colonnes dont les numéros sont consécutifs et dont le premier est X (ou 1, si X est omis). Cette instruction permet d'éviter la répétition de l'instruction précédente, notamment après chargement d'un fichier personnalisé.

5.8. - Instruction LEC (Sortie des valeurs d'une ligne pour plusieurs colonnes).

Elle s'écrit : LEC X [, Y, ..., Z]

Cette instruction a pour but d'imprimer sur la liste du terminal les valeurs correspondant à une région et contenues dans la colonne X (et les colonnes Y, ...Z, le cas échéant).

5.9. - Instruction EDI (Impression simple de colonnes).

Elle s'écrit : EDI X [, Y, ... Z]

Cette instruction entraîne l'impression sur la liste du terminal des valeurs contenues dans les colonnes X (et Y, ... Z). Le nombre maximum de colonnes autorisé est 6. L'impression a lieu sans titre, sans libellé.

Les langages de traitement et de recherche ainsi définis permettent toutes les opérations ordinaires menées par les utilisateurs de statistiques économiques localisées. Leur choix a été guidé par deux considérations :

1) Simplicité d'écriture, d'utilisation. En effet, les utilisateurs potentiels d'un tel système ne sont pas des spécialistes de la programmation. Il faut donc leur donner un langage aussi simple que possible et dont la description ne nécessite pas un méta-langage très complexe. La syntaxe du langage mis au point est de toute simplicité. Les codes choisis l'ont été pour des raisons mnémoniques sans que toutefois on puisse affirmer qu'il ne faille pas les modifier.

2) Simplicité opératoire. Il est relativement aisé et tentant, de construire des instructions permettant d'effectuer des traitements plus complexes, surtout dans le domaine statistique. Ce serait alors alourdir l'ensemble du sous-système de traitement sans être assuré que cela soit d'une grande utilité. Il a été jugé préférable de donner les moyens de se constituer un fichier d'interface dans lequel l'utilisateur pourrait stocker des informations extraites de la base de données et les traiter le cas échéant. Ces informations constituent alors le fichier d'entrée de son "modèle" qui est mis en oeuvre par un programme qui n'entre pas dans SAD II TS.

3.5.-Développement et Orientation de SAD II T.S.

Le système sad II TS évoluera comme le système 9:40 le permettra. En particulier : on modifiera les consultations de fichiers, en utilisant les ressource de l'accès direct. De plus, on utilisera le fichier table pour résoudre un problème de représentation interne. Les données des fichiers actifs sont représentées en entier et stockées dans des mots à 24 bits ; or il y a quelquefois des données qui, réduites avec le facteur d'échelle, dépassent $2^{23} - 1$. La valeur est alors le complément de $2^{23} - 1$ augmentée de 1. Pour éviter cet inconvénient rare (environ 1 série sur 1000 ou mieux une valeur sur 30 000 environ), on utiliserait la table pour renvoyer dans un fichier où les données en vrac seraient stockées sous forme réelle. Le pointeur pourrait aussi être utilisé pour donner une information sur la validité des opérations.

Le système sera élargi au cas départemental puis au cas départemental-régional. La structure générale ne devrait pas subir de profonds remaniements. La structure de la zone de travail sera modifiée.

SAD II TS devrait survivre comme système léger après l'apparition de SAD II. Celui-ci sera en mesure d'assumer la gestion de la base de données de SAD II TS. Ce dernier devrait se révéler pour certains utilisateurs peu spécialisés ou occasionnels comme d'un emploi plus économique que SAD II, mais ce serait au détriment de la puissance de SAD II dans le domaine de la recherche des séries disponibles notamment. De plus, les demandes de transfert d'informations entre les informations en archives externes ou internes et les fichiers actifs, ne seraient pas introduites dans le système . SAD II TS nécessiterait donc une intervention humaine pour la gestion de sa base de données par SAD II.

C O N C L U S I O N

Les banques de données surgissent, çà et là, très nombreuses, très diversifiées dans leurs objectifs, leurs méthodes et leurs performances. Elles répondent sans aucun doute à une nécessité et elles font naître, dans le domaine de l'informatique, de nouveaux problèmes, tant au point de vue technique qu'au point de vue humain.

L'objectif majeur d'une banque de données est de rapprocher l'utilisateur des fichiers de la base. Ce rapprochement doit pouvoir se faire dans le cadre de pensée et de travail de l'utilisateur. C'est pourquoi celui-ci n'a connaissance que des structures externes parmi toutes les structures mises en jeu ; en particulier, il n'a pas à connaître la structure interne de la base de données et celle élémentaire des fichiers ou même des articles. Il en résulte que l'un des facteurs du rapprochement de l'utilisateur des données est l'insertion d'une banque de données dans un ensemble plus vaste, un réseau de systèmes dans le cas de banques de données économiques, géographiques. Par exemple, ceci suppose que les transmissions de données puissent se faire facilement entre ordinateurs. Un autre facteur est également l'accès et la manipulation des données sur le mode conversationnel, avec toutes les nuances qu'implique l'application envisagée. Un second grand objectif est bien souvent la constitution d'un système d'archivage souple. Certains systèmes peuvent concilier, malgré l'importance des archives constituées, les deux objectifs. Un troisième objectif découle de l'aspect économique de l'utilisation. Il s'agit de faire partager au maximum d'utilisateurs, les avantages des structures introduites, soit en permettant des accès simultanés par des programmes d'interrogation, soit en donnant des moyens souples et faciles de programmer l'utilisation des données dans des programmes spéciaux.

Les méthodes de fonctionnement des banques de données restent très diversifiées selon les types d'applications pour lesquelles elles sont conçues. Elles restent aussi très liées aux configurations de calculateur employées, aux "operating system". La généralisation des mémoires à accès sélectif permet la création de structures très élaborées en vue de constamment améliorer les temps de réponse. L'apparition dans le software des "operating system", de modules permettant la constitution de banque de données entraînera aussi certaines évolutions. On pourra alors s'affranchir de certaines tâches, comme la gestion des terminaux, pour faire porter l'effort sur l'application elle-même.

Les performances sont évaluées bien souvent au niveau de l'exploitation et de l'interrogation en particulier. Cependant, c'est au niveau de la gestion que les problèmes, notamment dans les sciences humaines, vont se poser et souvent ils sont résolus de telle sorte que les performances globales du système sont diminuées. C'est surtout en fonction de la base de données que la banque évolue bien ou mal. La bonne évolution consiste à suivre celle des utilisateurs et à pouvoir s'adapter en souplesse aux nouveaux problèmes d'utilisation.

L'utilisation simultanée et partagée d'un très grand volume de données entraîne le développement d'une nouvelle activité dans l'informatique. L'activité banque de données vient s'ajouter aux activités traditionnelles : gestion, calcul scientifique et technique, contrôle de processus. Cette nouvelle activité intègre bien des techniques des précédentes.

En ce qui concerne les données manipulées par les sciences humaines, il convient de noter la difficulté d'utiliser rationnellement l'information disponible. Dans le domaine économique, malgré l'extraordinaire richesse de l'information, celle-ci est bien souvent utilisée pour la construction d'"indicateurs" synthétiques. La difficulté de mise en oeuvre de modèles statistiques et leur diffusion relativement restreinte font que ces indicateurs restent très primaires et ne reflètent en rien la finesse de l'information qui a permis de

de les élaborer.

C'est aussi sur le plan humain que l'utilisation des banques de données pose déjà des problèmes La formation indispensable à une bonne utilisation doit être fournie même si la banque de données est dotée de langages d'interrogation et de traitements de haut niveau.

Une banque de données n'est pas un système "à tout faire" ; c'est un système souple permettant l'accès à des données et des traitements , conçu pour des applications ou des classes d'applications définies.

REFERENCES BIBLIOGRAPHIQUES

1 - J.R. ABRIAL et autres
Le Projet SOCRATE I.M.A.G. Août 1970.

2 - J.M. MAESTRACCI Thèse (à paraître).
I.E.E.A.
Université des Sciences et Techniques - Lille

3 - Guide du Terminal et autres documents.
Publications "Telesysteme"

4 - "Software des Banques de Données".
Compte-rendu du colloque AFCET-IRIA.
Aix-en-Provence - Juin 1970.

5 - Le Système d'Accumulation des Données.
Faculté des Lettres et Sciences Humaines de Lille - 1968

UNIVERSITÉ DES SC. ET TECHN. LILLE
BIBLIOTHÈQUE UNIVERSITAIRE