Groupe de Travail I.N.S.E.E. - I.N.E.D. - ORSTOM

Manuel de dépouillement d'enquêtes

(version provisoire)

Avril 1977

PLAN de l'OUVRAGE

Chapitre 1

Introduction - Objectifs du manuel

Rédacteurs : L. BREAS,
JLB. BODIN

Etat de la rédaction : complète - provisoire

Plan du Chapitre :

- Objectifs du manuel,
- Plan du manuel,
- Organisation générale du dépouillement

CHAPITRE 1

INTRODUCTION

Le dépouillement des enquêtes statistiques relève, par nature, du traitement automatique de l'information : traitement clairement défini et en général assez simple à appliquer à un grand nombre de dossiers (ici des questionnaires) au stade du contrôle, de la correction et de la codification des données ; comptages nombreux, selon des critères très variés au stade de l'exploitation des résultats, etc...

On rencontrera lors de l'informatisation du dépouillement d'une enquête des problèmes de même nature que lors de l'informatisation d'un processus administratif avec une difficulté supplémentaire qui tient à ce que le système mis en place ne fonctionnera que pendant un laps de temps limité (sauf pour les enquêtes répétitives) alors qu'en gestion administrative les systèmes sont généralement destinés à fonctionner pendant plusieurs années, ce qui laisse le temps de mieux les étudier puis de les perfectionner si, à l'usage, il apparait qu'ils présentent des lacunes.

A quoi tiennent les difficultés ? Essentiellement au fait que vont être amenés à travailler ensemble des hommes dont la formation, les centres d'intérêts, les méthodes de travail, les langages techniques sont différents. Il faut craindre qu'ils se comprennent mal, que chacun enfermé dans sa spécialité ne prenne pas suffisamment conscience des objectifs et des contraintes de travail des autres.

Dans le cas du dépouillement des enquêtes, trois types d' "acteurs" sont en présence :

1. Les statisticiens,
2. Les agents chargés de la codification, de la saisie et de la gestion de l'enquête,
   (ci-après qualifiés de "gestionnaires"),
3. Les informaticiens.

.../...

(En principe, la distinction n'est pas toujours faite entre le statisticien et l'informaticien, soit qu'il s'agisse effectivement de la même personne, soit que le statisticien soit amené à empiéter sur le domaine de compétence de l'informaticien).

Le statisticien est l'initiateur, il définit l'enquête ou précise les objectifs et commande le travail. C'est pour l'essentiel un homme d'études, un chercheur hautement qualifié dont le souci est d'analyser les structures économiques et sociales ; l'enquête est pour lui un moyen d'y parvenir.

Les "gestionnaires" et l'informaticien sont des prestataires de service qui mettent au service du statisticien des techniques et des moyens humains et matériels propres à résoudre son problème. Techniques et moyens très différents de l'un à l'autre :

- pour la partie "gestion" de l'enquête, les moyens matériels sont assez frustes et les moyens humains importants ; la technicité est assez faible ; le travail est répétitif et simple ; le problème essentiel est de prévoir les charges, d'organiser les tâches, de les planifier, de les répartir entre les agents, de veiller au respect des consignes de travail (le nombre d'agents concernés est important), et au rendement de l'atelier,

- pour l'informaticien les moyens matériels sont très sophistiqués, les moyens humains peu nombreux, la technicité élevée avec une tendance à l'ésotérisme ; les tâches sont d'une grande complexité ; le délai de réalisations est souvent important ; la possibilité de les fractionner est limitée ; le contrôle de leur fiabilité est très délicat ; la prévision des charges est difficile et reste le plus souvent assez floue car elle dépend largement du niveau intellectuel et technique de ceux qui réalisent. L'informatique reste un "métier d'art" avec tous les impondérables que cela suppose.

Donc, trois types d'agents dont les profils sont très différents. Si le dialogue ne s'instaure pas, si chacun ne fait pas l'effort nécessaire pour acquérir une connaissance sommaire des techniques, des problèmes et des contraintes des deux autres, le risque de conflit, voire d'échec est sérieux. Par exemple, il est tout à fait nécessaire que :

- le statisticien admette que la présentation du questionnaire peut avoir

.../...

des incidences importantes sur le travail du gestionnaire et de l'informaticien : qu'il accepte, autant que faire se peut, les modifications que ceux-ci lui proposeront afin de faciliter leurs tâches ; qu'il admette aussi qu'un système informatique est une construction délicate que des modifications, qui lui paraissent de détail, peuvent compromettre très sérieusement.

- l'informaticien prenne conscience de ce que la logique informatique est souvent bien loin du sens commun et qu'il lui faudra apprendre, s'il souhaite être compris, à décoder l'information qu'il transmettra à ses partenaires sous forme de notes, d'états informatiques, de messages d'anomalies , qu'il se pénètre du problème à traiter afin d'être en mesure de juger de la pertinence de ce que l'on lui demande, et de proposer des solutions alternatives qui permettent d'aboutir au même résultat à moindre coût.

Le but principal que se sont fixés les auteurs de ce manuel est de faciliter l'ouverture du dialogue entre le statisticien et l'informaticien en faisant le point des problèmes et des méthodes de dépouillement d'enquêtes. On n'y trouvera rien d'original, ni de révolutionnaire. Il s'agit seulement d'un point, sans aucun doute incomplet, des acquis (de l'I.N.S.E.E. pour l'essentiel) en la matière. Un informaticien rompu au traitement des enquêtes statistiques trouvera sans doute les chapitres relatifs au traitement sur ordinateur bien banals. Un praticien de la statistique aura probablement la même réaction devant les chapitres qui traitent de contrôle et de codification.... L'important est qu'il décrive l'ensemble du processus de dépouillement d'enquête, ce qui, à notre connaissance, n'avait jamais été fait. Certains chapitres s'adressent d'abord aux statisticiens, d'autres aux gestionnaires, d'autres encore aux informaticiens. Il nous parait essentiel qu'il soit lu entièrement par tous afin que chacun ait une vue globale du processus et acquiert cette connaissance minimale des techniques et problèmes de ses partenaires qui nous parait indispensable à un travail d'équipe fructueux.

Un aspect du travail a été pratiquement passé sous silence dans ce manuel : c'est celui des opérations de collecte sur le terrain (par enquête). En effet, d'une part il existe de nombreux manuels traitant de cet aspect du problème, et d'autre part, il y a une certaine indépendance entre les opérations de collecte et les opérations de dépouillement d'enquête qui font l'objet de ce manuel. Le point de rencontre entre ces deux types de préoccupations est le questionnaire qui

.../...

doit être adapté aux impératifs de la collecte (progression des questions, libellé, présentation, etc...) comme à ceux du dépouillement ; le statisticien sera le "contact" avec l'homme du terrain, mais très généralement, ce sera lui le responsable de la collecte.

PRESENTATION GENERALE DU MANUEL

1. La dernière partie du présent chapitre décrit les différentes phases de la conception, de la réalisation et du dépouillement d'une enquête. Elle en présente sommairement le contenu et décrit les liens fonctionnels qui les unissent.

2. Les chapitres 2 (contrôles et redressements) et 3 (codifications) présentent les principaux types de contrôles, redressements et codifications utilisés dans le traitement des enquêtes statistiques [Point de vue du statisticien.] C'est un sujet qui peut être, et a été (références bibliographiques ?), plus largement développé qu'il ne l'a été ici. On s'est volontairement limité à une description relativement sommaire destinée à servir d'aide-mémoire au statisticien et de présentation du problème au "gestionnaire" et à l'informaticien. Pour ce dernier notamment, lorsqu'il est habitué à traiter des problèmes de gestion administrative, il était nécessaire de présenter les caractères très spécifiques des contrôles et corrections statistiques.

3. Les chapitres 4 & 5 sont consacrés à l'organisation de l'atelier de traitement manuel et à la saisie des données. C'est là le domaine de compétence propre des "gestionnaires".

On s'est attaché à mettre en évidence les interactions qui existent entre cette phase du traitement et la phase proprement informatique. En effet, l'apport des techniques informatiques déborde de plus en plus du domaine traditionnel des contrôles "batch" et de la tabulation déjà pris en compte par la mécanographie classique. Il convient donc d'organiser les travaux des "gestionnaires" en tenant compte de ces nouvelles possibilités : messages d'anomalies expédiés par le centre de traitement informatique à l'atelier "manuel", nouvelles possibilités offertes par les méthodes modernes de saisie etc...

.../...

4. Le chapitre 6 (forme et contenu du questionnaire) est un chapitre pivot entre le traitement manuel et le traitement informatique$_q$. Il n'a pas pour objet de décrire la ou les méthodes de construction d'un questionnaire, ce qui a souvent été fait par ailleurs (références bibliographiques ?) mais seulement de présenter les aménagements de mise en forme, présentation, précodification, préimpression de données d'identifications issues du fichier de lancement, etc... qu'il peut être souhaitable d'apporter au questionnaire afin de faciliter et simplifier les tâches des gestionnaires et des informaticiens, accrôître la fiabilité des contrôles, etc...

5. Les chapitres 7 (organisation des fichiers séquentiels), (8 (la chaine d'apurement) et 9 (la tabulation) présentent les traitements informatiques classiques. Ils forment le noyau central de ce manuel.

5.1. Le chapitre 7 présente sommairement la notion de fichier. Son principal objectif est de mettre en valeur l'interaction entre structure des fichiers et la structure des programmes ; si la structure des fichiers est définie par des règles cohérentes et générales, il en ira de même de la structure des programmes dont la construction se trouvera ainsi facilitée. C'est pourquoi on s'est limité à la description d'un seul type de structure qui se révèle particulièrement exemplaire.

5.2. Le chapitre 8 décrit la chaine d'apurement c'est-à-dire l'ensemble des opérations qui, à partir des données, sur cartes ou support magnétique, en provenance de la saisie concourent à l'élaboration d'un fichier exhaustif et apuré de ses erreurs, c'est-à-dire susceptible d'être exploité statistiquement. La démarche consiste à mettre en valeur le fait que le nombre de ces opérations est limité et connu, que certaines posent des problèmes types auxquels on peut appliquer des solutions-types. Une fois tout ceci recensé, il apparait que construire une chaine d'apurement consiste à assembler logiquement un ensemble de modules, dont le nombre varie selon les caractéristiques de l'enquête, dont certains sont entièrement standards d'une enquête à l'autre, d'autres sont partiellement standards, d'autres encore entièrement spécifiques.

.../...

5.3. Le chapitre 9 traite de la fabrication des tableaux statistiques. L'objectif est double :

a) présenter la notion de tableau ou décrire les concepts, proposer un vocabulaire sans ambiguités qui puisse être utilisé par les statisticiens comme par les informaticiens. Cet aspect est essentiel : les statisticiens savent, certes, ce qu'est un tableau statistique, mais cette connaissance acquise, un peu sur le tas, au fil de l'apprentissage de leur métier, est, bien souvent, mal formalisée, sans soubassement théorique ; il en résulte une certaine difficulté à l'exposer de manière complète et cohérente. Quant aux informaticiens non habitués au traitement des enquêtes statistiques, il s'en font une idée limitée à ce qu'ils ont eu l'occasion de réaliser lors d'applications administratives, c'est-à-dire des états très simples, le plus souvent à une seule dimension, très loin de la complexité que peuvent atteindre les tableaux utilisés par des statisticiens.

b) de la même manière que dans le chapitre précédent, faire apparaitre que la fabrication des tableaux repose sur la mise en oeuvre d'un ensemble bien défini d'opérations qu'il suffit d'agencer correctement pour élaborer des programmes de tabulation qui répondent à des règles de construction bien précises.

6. Le chapitre 10 (l'information en sortie) traite de problèmes qui sont à la frontière de l'informatique et de l'imprimerie et de la diffusion. L'information produite lors de l'exploitation des résultats devra être diffusée, reproduite, archivée,... Le chapitre recense les méthodes et les moyens dont on peut disposer pour ce faire. Il est rare que l'informaticien et le statisticien se préoccupent des problèmes matériels de diffusion. Il est vrai que l'importance de ces problèmes est liée à l'importance du volume de "sorties" envisagées. Il n'en est pas moins apparu nécessaire aux auteurs de consacrer un chapitre à ce sujet.

7. Le chapitre 11 (cahier des charges et documentation) présente l'instrument de communication entre les parties associées au dépouillement de l'enquête (statisticiens, "gestionnaires", informaticiens). Nous présentons dans ce chapitre une solution, le cahier des charges, qui doit, avant tout, permettre le dialogue entre ces partenaires. Il faut que ce dialogue soit clair, sans ambiguités ; il faut que soient clairement précisées les tâches à réaliser, les charges de

.../...

travail qu'elles induisent, les délais de réalisation ; que ce soit entre les équipes (statistiques, gestion manuelle, informatique), ou au sein des équipes. Il faut donc qu'il y en ait une trace écrite à laquelle on pourra se référer en cas de litige. Le chapitre 11 propose un schéma d'organisation du cahier des charges dont le but est de préciser les responsabilités de chacun, l'affectation des tâches, les charges et délais de réalisation, indique comment ses différentes parties pourront s'élaborer au fil de l'avancement des travaux, décrit le contenu des différents dossiers (d'analyse, de programmation, d'exploitation...) qui viendront le compléter. Le schéma proposé n'est certes pas unique ; disons simplement qu'il a concrètement fait la preuve de son efficacité dans un certain nombre d'applications.

8. On trouvera enfin, en annexe, une présentation commentée de quelques logiciels de dépouillement d'enquêtes. Là non plus on n'a pas cherché l'exhaustivité. On s'est limité à 3 exemples, connus des auteurs du manuel, qui sont représentatifs de 3 grandes classes de logiciels :

- Le 1er est très puissant mais requiert des moyens machines importants (LEDA) - 256 k-octets de mémoire,

- Le 2ème est plus juste mais peut fonctionner sur des machines de faible capacité (GROULT + CENTS),

- Le 3ème enfin est un système ouvert dont les fonctions de base sont relativement réduites mais conçu pour intégrer aisément des fonctions complémentaires (qu'il faut programmer en FORTRAN) propres à telle ou telle application (SIXE - INED).

ORGANISATION GENERALE DU DEPOUILLEMENT

L'organisation générale du dépouillement consiste à faire un descriptif global des différentes phases et de leurs liaisons fonctionnelles, à définir le "chemin critique" c'est-à-dire, notamment, à repérer les phases pour lesquelles un retard de réalisation entrainera ipso-facto un retard pour le travail dans son ensemble. La mise au point d'un "graphe des contraintes" faisant apparaitre les interactions et la durée prévisible de chaque phase pourra aider à mettre au point le shhéma général du dépouillement de l'enquête.

La liste des phases indiquée ci-après est donnée à titre purement indicatif. Il conviendra de l'établir pour chaque nouveau travail.

A - Définition des objectifs

- A0 - Objectifs généraux,
- A1 - Ebauche du plan de tabulation,
- A2 - CHoix des questions à poser,
- A3 - Choix des codes et nomenclatures utilisés pour le chiffrement et la saisie,
- A4 - Plan d'échantillonnage - Effectif enquêté,
- A5 - Périodicité.

B - Analyse des moyens disponibles et souhaitables

- B1 - Moyens humains (personnels de chiffrement, de gestion, de saisie des données, d'analyse-programmation,...),
- B2 - Moyens matériels (saisie, traitement,...)

C - Définition des contrôles et redressements

- C1 - Avant la saisie (contrôles "manuels"),
- C2 - Pendant la saisie,
- C3 - Sur ordinateur,
- C4 - Recours au redressement manuel

D - Choix du mode de saisie

.../...

E - Choix du mode de collecte (par enquêteur, par voie postale, par sous-produit d'opérations administratives, ...)

F - Définition du questionnaire,

G - Déroulement de l'enquête "sur le terrain" (collecte),

NB - En fait, cette phase peut se décomposer en un grand nombre de sous-phases élémentaires. De nombreux ouvrages ayant été consacrés à la collecte, on ne cherchera pas à détailler la présentation de cette phase.

H - Organisation de la saisie et de l'apurement du fichier

- H1 - Organisation et définition des tâches de l'atelier de chiffrement,
- H2 - Organisation et définition des tâches de l'atelier de saisie et codage des formats de saisie,
- H3 - Analyse et programmation des traitements sur ordinateur,
- H4 - Organisation de l'exploitation sur ordinateur,
- H5 - Définition des liens fonctionnels entre atelier de chiffrement, atelier de saisie et ordinateur,

I - Exploitation des résultats

- I1 - Plan de tabulation définitif,
- I2 - Définition des codifications,
- I3 - Analyse et programmation des codifications,
- I4 - Analyse et programmation des tableaux,

J - Analyse des résultats.

La phase A est de la responsabilité du statisticien qui devra à ce stade :

- définir les objectifs généraux de l'enquête et montrer en quoi elle peut satisfaire les besoins exprimés,

- rédiger un avant-projet précisant la liste des questions et les tableaux qu'on espère élaborer, le nombre d'unités à interroger, le mode d'enquête, etc...

- essayer de préciser les éléments de coût,

- soumettre le projet aux différentes administrations ou aux différents organismes intéressés, par le canal éventuel des commissions ou comités prévus le cas échéant à cet effet (comme, par exemple, en France, le Conseil National de la Statistique).

.../...

La phase B est également de la responsabilité du statisticien, mais il aura besoin du concours des techniciens-informaticiens ou autres - pour améliorer ses estimations et évaluations. Dans cette phase, interviennent aussi les responsables administratifs et financiers de l'organisme dans lequel travaille le statisticien. En fait, pendant toute la préparation de l'enquête, il y aura "feed-back" entre les phases opérationnelles (telles que C, D, H, I, etc...) et cette phase B, dans la mesure où il s'avèrera nécessaire d'ajuster les choix techniques aux moyens humains et financiers disponibles.

Seules certaines des phases C à J font l'objet de ce manuel ainsi qu'il a été déjà exposé. Les phases D, E, et F, ne seront abordées qu'en ce qui concerne leurs interactions avec les autres phases du dépouillement. Les phases G et J ne seront pas abordées.

G

On trouvera ci-après, à titre purement illustratif, un exemple d'analyse sous forme de tableau des interactions entre les différentes phases et sous-phases, ainsi que le graphe des contraintes qu'on peut déduire de ce tableau.

MANUEL DE DEPOUILLEMENT D'ENQUETES

F. PRADEL de LAMAZE

Chapitre II : Contrôles et Corrections d'Erreurs

Une enquête statistique présentera inévitablement un certain nombre d'erreurs, les unes fondamentales et d'autres négligeables. Il importe au plus haut point d'éviter les erreurs fondamentales qui mettent en cause la validité même de l'enquête et son intérêt, mais ce serait une grave faute de négliger au départ les erreurs minimes. Si, en effet, de petites erreurs n'ont pas, du point de vue statistique, d'importance capitale, elles entraîneront tout au long de la chaîne d'exploitation, y compris dans les tableaux de résultats, des déboires considérables allant de certaines incohérences à des rejets purs et simples par les chaînes informatiques. L'opération de contrôle et éventuellement de correction des erreurs doit donc être étudiée avec beaucoup de soins, en vue de plusieurs objectifs.

Du strict point de vue de l'exploitation d'une enquête, qui nous retiendra dans cet ouvrage, le but des contrôles est relativement clair : il s'agit d'obtenir un fichier tel que la production de tableaux se fasse sans difficultés. Cela se traduit par quelques notions simples : aucun enregistrement ne doit manquer (ou à fortiori se trouver en trop !), chaque enregistrement doit être complet, toutes les modalités d'un caractère rencontrées dans un enregistrement (y compris 0, non déclaré, ne sait pas, etc..) doivent être prévues et avoir leur place dans les tableaux des résultats.

Au statisticien toutefois, d'autres objectifs s'imposent ; un fichier informatiquement parfait risque, en effet, de recouvrir une enquête déplorable et il importe évidemment d'éviter ce risque, d'autant plus grave que l'informatique permet la production rapide d'une très grande quantité de tableaux dont on n'a pas toujours le loisir d'examiner la valeur avant de les diffuser. Les responsables d'enquêtes devront donc avoir souci d'utiliser les ressources de l'informatique pour les aider à préciser certains points.

L'important sera en particulier de détecter le plus rapidement possible les erreurs les plus graves, et de savoir en tirer les conséquences même les plus lourdes: retour sur le terrain, élimination de certaines parties du questionnaire, restrictions sur la diffusion des résultats, etc.. Il est, en effet, presque toujours possible de donner une "propreté" formelle à un fichier, mais, aussi bien exploitée, une mauvaise enquête ne saurait devenir bonne par la magie de l'exploitation informatique.

En dehors des cas, heureusement assez rares, ou la validité même de l'enquête peut être remise en cause, l'objectif du statisticien sera pour l'essentiel :

... /...

- d'éliminer des erreurs de détail sans intérêt statistique mais souvent gênants ;

- d'assurer la cohérence des résultats, en particulier d'un tableau à l'autre. Ce dernier point revêt une certaine importance et amène à fournir des tableaux d'une grande précision (souvent 7 ou 8 chiffres significatifs) comptable. On doit tenir compte de l'utilisateur qui sera souvent un peu perdu par divers concepts, en particulier le champ sur lequel porte un tableau, et pour lequel l'exactitude comptable peut servir de point de repère, voire de bouée de sauvetage.

Cet ouvrage s'intéressant essentiellement à l'exploitation informatique des enquêtes, nous n'insisterons pas davantage sur la nécessité de contrôles très serrés sur le terrain, contrôles pour lesquels l'intervention de l'informatique est à priori assez faible. Mais il nous paraît essentiel de rappeler au début de ce chapitre, et de ne jamais le perdre de vue, qu'une enquête se passe d'abord et avant tout sur le terrain, et que les erreurs initiales se retrouveront, peut être trop souvent masquées, dans les résultats.

Par ailleurs, il est toujours possible, souvent souhaitables, de décompter les erreurs et les corrections apportées au cours de l'exploitation. Mais il serait illusoire de penser que toutes les erreurs soient pour autant éliminées, ni à fortiori décomptées. On doit toujours savoir au mieux ce que l'on a fait en ce sens (nombre d'erreurs repérées, méthode de correction, biais éventuel indroduit par ces corrections), mais ne pas se leurer à l'excès. La seule indication réelle que donneront des tableaux de décomptes d'erreurs est subjective : si un bon filtrage ne laisse apparaître que peu d'erreurs, l'enquête est probablement bonne, s'il laisse apparaître beaucoup d'erreurs on doit être très prudent.

A - Contrôles des enregistrements.

Les données de l'enquête étant enregistrées sur support magnétique, on devra procéder à deux types de contrôles : contrôle de la forme des enregistrements et contrôles logiques des liaisons internes à ces enregistrements ; éventuellement si le fichier est organisé selon une règle préétablie (fichiers hiérarchisés, fichiers chronologiques, etc..) on procèdera de plus à un contrôle de structure. Ces contrôles se feront bien entendu selon un ordre défini, à priori le suivant : contrôles de structure, contrôles de forme, contrôles logiques. Mais en fait pour chaque enquête et éventuellement chaque type de contrôle, cet ordre peut être changé. Le contrôle n'étant pas une fin en soi mais ayant pour but essentiel la correction des erreurs, par un procédé quelconque (correction manuelle par retour au dossier, correction automatique ou aléatoire, etc..), on devra aussi effectuer ces corrections selon un ordre déterminé, généralement le même que celui adopté pour les contrôles. La dimension du fichier aura une grande importance dans le choix de la

... /...

stratégie, chaque programme de détection et de correction d'erreur nécessitant un passage de la totalité du fichier ; sur de petits fichiers on pourra donc envisager plusieurs programmes successifs, et même éventuellement des corrections par itérations : sur de gros fichiers par contre il y aura intérêt à utiliser un programme unique, permettant au cours du déroulement du fichier de procéder à toutes les opérations de contrôle et de corrections ; un tel programme est évidemment beaucoup plus difficile à mettre au point qu'une succession de programmes légers, et parfois, mais pas toujours, moins satisfaisant. En tout état de cause plus le fichier sera important meilleures pourront être les corrections apportées, non au niveau individuel mais dans leur ensemble.

1 - a) Contrôles de forme :

Un fichier d'enquête se présente comme une suite d'enregistrements, chaque enregistrement correspondant à un questionnaire. Nous verrons plus tard comment doit être contrôlée la structure du fichier.

1 - a.a) Exhaustivité :

Le premier contrôle doit porter sur l'exhaustivité du nombre d'enregistrements. Généralement ce nombre est connu et il suffira de décompter les enregistrements.

Comme en cas de non coïncidence on ne saurait pas quels sont les enregistrements manquants, ou excédentaires, on devra avoir avoir pris soin de numéroter les questionnaires soit selon un ordre naturellement ascendant (de 1 à n) soit selon des séquences successives (de 1 à n $<$ 100, puis de 101 à n' $<$ 200, de 201 à n" $<$ 300 etc.. par exemple). Dans le premier cas on vérifie la présence, une seule fois, de chaque questionnaire, c'est-à-dire de tous les numéros compris entre 1 et n (à noter qu'il n'est pas nécessaire pour effectuer cette vérification que le fichier soit en ordre), dans le second cas on vérifie pour chaque séquence la présence des numéros successifs (la connaissance des nombres n, n', n"... est alors nécessaire). Une troisième méthode enfin consiste à comparer les enregistrements numérotés d'une façon quelconque à une liste des questionnaires établie par ailleurs.

Exemples :

Premier cas : on a interrogé 10 223 personnes et pour chacune a été établi un questionnaire numéroté de 1 à 10 223. On vérifie simplement qu'à chaque nombre inférieur ou égal à 10 223 correspond 1 enregistrement et un seul.

Deuxième cas : on a interrogé un certain nombre de personnes dans 50 communes. On numérotera les questionnaires en deux temps : un numéro de commune (de 01 à 50) et un nombre d'ordre dans la commune (de 001 à n). Sachant quel est le nombre n dans chaque commune, on vérifie la présence d'un questionnaire pour chacun des numéros à cinq chiffres possibles (01001 à 01253, 02001 à 02521 etc.. par exemple).

.../...

Troisième cas : on a interrogé 10 000 personnes caractérisées par un numéro d'identité à 10 chiffres. On confrontera alors la liste des numéros de personnes interrogées à celle des enregistrements.

De tels contrôles, extrêmement simples, permettent d'éviter souvent de très graves erreurs : lots de questionnaires oubliés (voire saisies deux fois), bandes magnétiques incomplètes, enregistrement erroné non effacé, etc.. La correction est alors évidente.

Nota : Il arrive assez souvent, dans le cas correspondant aux exemples 1 et 2, que les nombres n, n', etc.. ne soient pas connus à priori. Dans la mesure où l'on a adopté une numérotation séquentielle, cela ne sera pas trop grave, le numéro le plus élevé rencontré dans les enregistrements fournissant à postériori ce nombre n, les "trous" dans la séquence restent évidents. Le seul risque est alors que ce soient les derniers numéros qui manquent, mais lorsqu'il ne manque que quelques enregistrements il est rare que ce soient les derniers, lorsqu'il en manque un lot important on a toujours une indication du nombre approximatif de questionnaires,

1 - a.b.) Identifiants

Chaque enregistrement doit être repérable par un numéro. Dans les cas simples un numéro d'ordre suffira (voir 1-a.a), dans d'autres cas plus complexes il sera judicieux d'identifier l'enregistrement par un numéro analytique. Par exemple, dans un questionnaire auprès des ménages chaque individu sera identifié par son numéro propre, son numéro de ménage et éventuellement son appartenance à une unité géographique donnée. L'exactitude de ces identifiants est d'une très grande importance car c'est sur eux que se feront certains rapprochements (attribution d'un salaire à un individu, puis à un ménage, rapprochement des dates de naissances des enfants avec celle de leur mère, etc..). Une vérification des identifiants est possible si ceux-ci ont été construits de façon à obéir à certaines lois relativement simples, souvent appelées "clés de contrôle".

Exemple : on identifiera un individu par son numéro de commune de résidence, son numéro de ménage à l'enquête et son numéro d'ordre à l'intérieur du ménage. A cet identifiant comprenant 8 chiffres on ajoutera systématiquement un 9ème chiffre tel par exemple que la somme des chiffres constituant l'identifiant soit paire :

205 623 01 , la somme est impaire, on adjoint 1
205 623 02 , la somme est paire, on adjoint 2

Les identifiants complets deviennent :

205 623 01 1 et 205 623 02 2

En vérifiant la parité on voit qu'on détectera immédiatement les identifiants à somme impaire donc erronés.

.../...

La "clé" prise comme exemple est d'une grande simplicité, mais malheureusement de ce fait insuffisante. Une erreur transformant un 3 en 5 par exemple passera inaperçue. Selon l'importance attachée à l'identifiant on devra donc rechercher des clés plus ou moins complexes, une grande simplicité rendant la clé assez fragile, une grande complexite la rendant plus difficile à établir et surtout à calculer. Un peu plus délicate à manier que la précédente, mais beaucoup plus sûre serait par exemple la condition "la somme des chiffres de l'identifiant doit être divisible par 7". On a alors :

205 623 01 2
205 623 02 1
205 623 03 0
205 623 04 6

1 - a.c) Valeurs impossibles. :

En examinant une par une les réponses enregistrées on détectera aisément certaines impossibilités, soit naturelles (un sexe ne saurait être que masculin ou féminin), soit imposées par l'enquête (une enquête sur des élèves d'une école de garçons ne saurait comprendre de filles). L'examen question par question est laborieux mais les moyens informatiques permettent de la réaliser très rapidement ; sans préjuger de l'attitude à adopter en cas de découverte d'une erreur, on s'efforcera de détecter ces erreurs en établissant une liste de tous les cas possibles. Parmi ces cas possibles certaines éventualités sont parfois à envisager bien qu'elles paraissent à priori invraissemblables, par exemple pour le sexe on peut avoir trois cas : masculin, féminin et non déclaré, ce troisième cas étant distinct d'une déclaration erronée (traduit en code on pourra ainsi accepter un sexe = 1 (masculin) = 2 (féminin) X (non déclaré), mais on refusera tout code différent (0, 3, 4, ... Z ...) ; pour un salaire on doit vérifier qu'il soit situé dans une certaine fourchette, mais on peut accepter qu'il soit nul (si par exemple l'enquête ne relève que les salaires reçus au cours d'une période donnée, etc..).

D'une façon générale il importe de bien distinguer parmi les éventualités "anormales", celles qui correspondent à une absence de réponse (non déclaré, refus de réponse) et celles qui sont effectivement nulles. Un zéro réel est souvent acceptable tel quel alors qu'un "non réponse" ne correspond que rarement à un état réel nul.

Cette opération de détection des "dépassements de code" présente un grand intérêt mais aussi un certain danger. C'est, en effet, lors de cette étape que l'on va détecter le plus grand nombre d'erreurs apparentes, et l'on s'efforcera ensuite de les corriger. Mais en fait ne sont détectées que les erreurs souvant minimes, surtout lorsque les éventualités acceptables sont nombreuses. Les cas "hors code" proviennent plus souvent d'erreurs de perforation, relativement rares, que d'er-

.../...

reurs de chiffrement, beaucoup plus fréquentes : une difficulté lors du chiffrement est en effet rarement résolue par une codification "hors norme" mais le plus souvent par le recours à une codification "fourre tout" (autres cas, non déclaré,...) voire par une codification arbitraire mais formellement acceptable.

1 - b) Contrôles de structures :

Les contrôles formels dont nous avons parlé dans les paragraphes précédents peuvent s'effectuer sur un fichier en désordre, dès lors que sur ce fichier chaque enregistrement est bien individualisé. Mais le plus généralement on aura tout intérêt à travailler sur des fichiers ordonnés, selon des règles simples ou plus élaborées. (A noter que les contrôles de structure peuvent selon la nécessité précéder ou suivre les contrôles formels, mais généralement il vaut mieux qu'ils précèdent).

1 - b.a) Fichiers ordonnés simples :

Nous appelerons fichier ordonné simple un fichier dans lequel les enregistrements sont distincts les uns des autres et se suivent selon une loi simple : ordre chronologique, ordre ascendant des numéros de questionnaire, ordre géographique,etc.. De tels fichiers présente de l'intérêt dans la mesure où ils permettent d'isoler très simplement une portion du fichier (données portant sur telle année, sur telle unité géographique, etc..). Leur contrôle est relativement simple, et peut aisément être couplé avec le contrôle d'exhaustivité : l'ordre étant défini (par exemple classement des enregistrements selon l'ordre alphabétique des communes dans des régions) on classe le fichier selon cet ordre et on vérifie que chaque enregistrement se trouve bien à sa place. Lors de ces opérations on trouvera souvent, soit rejetée en fin du fichier soit isolés au coeur du fichier quelques enregistrements inclassables. Il pourra s'agir simplement, dans le cas le plus fréquent, d'erreurs d'identification (numéros mal perforés, erreur sur la commune de résidence, etc..) ; parfois on détectera des erreurs plus graves : absence complète d'une année (confusion entre la date de l'enquête et celle de l'évènement observé par exemple), présence de deux lots de questionnaires sur une même commune (confusion par exemple entre lieu de résidence et lieu de travail, ou erreur systématique sur le chiffrement de la région, etc ..).

1 - b.b.) Fichiers hiérarchisés :

Il est souvent commode de classer les fichiers en utilisant les "facteurs communs". Par exemple, lors d'une enquête démographique les individus seront observés à travers les ménages et dans une commune de résidence. Les caractères propres à l'individu lui sont personnels, mais il partage avec d'autres individus certains caractères communs (appartenance à un ménage de n personnes, résidence dans un logement sans eau, résidence dans une commune de n habitants, etc..). Plutôt que de répéter dans chaque enregistrement individuel des caractéristiques communes on

... /...

préfèrera souvent enregistrer séparément les données individuelles et les données collectives, et organiser le fichier de façon à pouvoir réattribuer à chaque individu ses caractères collectifs. En conservant l'exemple précédent on pourra organiser un fichier à trois niveaux hiérarchiques :

niveau 1 : commune de résidence
niveau 2 : ménage
niveau 3 : individu

Les enregistrements de niveau 1 comprendront toutes les caractéristiques utiles de la commune : caractère urbain ou rural, taille, situation géographique, présence d'une école, d'un hôpital, etc...

Les enregistrements de niveau 2 porteront sur les caractéristiques du ménage : nombre de personnes (éventuellement par sexes), âge du chef de ménage, profession et catégorie socio-professionnelle du chef de ménage, nombre d'actifs, caractéristiques de l'habitat, etc...

Les enregistrements de niveau 3 enfin ne concerneront que les caractères propres à l'individu : sexe, âge, état matrimonial, revenu individuel, etc...

On voit que, à condition de disposer d'un identifiant commun parfaitement sûr, il est loisible de rapprocher les caractéristiques d'un individu de celles de sa commune de résidence, même dans le cas où les enregistrements de niveau 1 sont "physiquement" éloignés de ceux de niveau 3. On préfère toutefois le plus souvent fondre les 3 niveaux selon une règle simple :

A la suite du premier enregistrement de niveau 1 (commune de résidence) on range le premier enregistrement de niveau 2 (ménage résidant dans cette commune) puis les enregistrements de niveau 3 (individus appartenant à ce ménage) correspondants, ensuite le second enregistrement de niveau 2 suivi des enregistrements de niveau 3 correspondants, etc... Ainsi à la suite de chaque ménage sont rangés tous les individus lui appartenant, à la suite de chaque commune tous les ménages et tous les individus y résidant.

L'adoption d'une telle hiérarchisation des fichiers présentera des avantages considérables pour la confection des tableaux (plus grande rapidité, facilité des tris, cohérence des résultats,etc..) mais nécessitera un contrôle très étroit de la structure du fichier, toute erreur entraînant une rupture de l'équilibre recherché.

Le contrôle devra reposer sur des principes aussi simple que possible. Dans l'exemple choisi ces principes sont les suivants :
- aucun enregistrement de niveau 3 ne peut exister sans être précédé des niveaux 2 et 1 auxquels il doit se rattacher,
- tout enregistrement de niveau 2 doit être suivi d'au moins 1 enregistrement de niveau 3 (il n'existe pas de "ménage" vide, mais il peut exister des ménages d'une seule personne),

.../...

- tout enregistrement de niveau 1 doit être (sauf si explicitement on a prévu l'existence de commune sans population enquêtée) suivi d'au moins 1 enregistrement de niveau 2 et d'au moins 1 enregistrement de niveau 3.

A ces principes de base on pourra adjoindre, ce qui peut permettre de détecter certaines erreurs, des principes complémentaires : ordre dans la numérotation des ménages (niveau 2) et des individus (niveau 3), nombre maximum des ménages par communes et d'individus par ménage, etc...

1-b.c.) Rapprochement des fichiers:

Proche dans son principe du cas précédent, le rapprochement de 2 fichiers à priori indépendants pose certains problèmes complémentaires. Supposons par exemple que l'on dispose d'une part d'un fichier concernant des salariés d'une entreprise (sexe, âge, niveau professionnel, etc...) d'autre part, d'un fichier donnant les salaires de ces employés. Le rapprochement n'est évidemment possible que si l'on dispose sur l'un et l'autre fichier d'un identifiant commun et sûr (nom de l'employé par exemple) (cet identifiant n'est pas toujours suffisamment sûr, les cas d'homononymies étant relativement fréquents). Mais cette condition n'est pas suffisante, il faut de plus qu'à chaque individu figurant sur le fichier 1 corresponde un enregistrement sur le fichier 2 et réciproquement. Il sera donc nécessaire pour permettre ce rapprochement de vérifier que cette concordance existe et sinon de la créer artificiellement (soit par élimination des individus ne figurant que sur l'un des fichiers, soit par création, éventuellement grâce à un complément d'enquête, d'un fichier supplémentaire).

Dans le même ordre d'idée le rapprochement de 2 fichiers de population à des dates différentes pourra nécessiter la création d'enregistrements concernant des individus fictifs (à naître ou décédés) afin d'éviter la recherche indéfinie d'un individu disparu du fichier pour une cause connue (décès), évidemment distincte de celle d'un individu disparu par erreur.

1- c.) Contrôles logiques:

Tous les types de contrôles envisagés jusqu'ici ne concernaient que la forme du fichier, il s'agissait de s'assurer que rien ne viendrait troubler la production de résultats formellement acceptables. Dans la mesure du possible les contrôles doivent être poussés davantage, la cohérence de forme risquant de cacher des incohérences de fond.

1 - c.a.) Cohérence interne:

A l'intérieur d'un questionnaire les questions sont rarement indépendantes, et le rapprochement des réponses pourra permettre soit d'éliminer des erreurs de détail soit éventuellement de détecter des questionnaires inexploitables (réponses volontairement incohérentes, incertitudes trop nombreuses, etc...).

.../...

Il sera d'ailleurs nécessaire lors de la rédaction même du questionnaire de prévoir certaines questions redondantes ou apparemment inutiles, dont le but est précisément de vérifier la qualité des réponses. Ce rapprochement est une opération extrêmement délicate, car souvent les contradictions n'apparaissent que par la confrontation de 3 ou 4 réponses, voire davantage. Il importera donc de limiter ces contrôles, d'une part sur les incohérences simples et évidentes, d'autre part sur les questions fondamentales.

Rarement les incohérences sont rigoureusement inacceptables (du type avoir 15 ans et avoir eu 5 enfants, habiter une maison sans eau courante et avoir une douche, etc..) mais le plus souvent elles se situent dans une marge plus ou moins large d'acceptabilité (avoir 18 ans et 5 enfants, être chef de ménage et n'avoir aucun revenu, ..).

Nous pensons qu'en fait ces problèmes doivent être résolus sur le terrain, à la rigueur lors des contrôles en bureau mais non au moment de l'exploitation. Le seul rôle, au demeurant important, d'une recherche d'incompatibilités internes sur les fichiers déjà enregistrés doit se limiter à notre avis à l'élimination d'erreurs matérielles (erreurs de chiffrement, erreurs d'unités), on a une aide aux contrôles en bureau (liste de questionnaires suspects, de cas particuliers trop nombreux, etc..).

Généralement donc on examinera question par question les incohérences inadmissibles d'une part, les incohérences fortement suspectes d'autre part, par référence à un ou deux critères choisis comme déterminants. Si, par exemple, on étudie le nombre de naissances survenues dans l'année pour une femme, le critère d'âge pourra faire ressortir comme situation impossible une naissance pour une mère d'âge inférieur à 10 ans ou supérieur à 60 ans, comme suspectes les naissances parmi les mères d'âge 10, 11 ou 12 ans ou 50 à 59 ans, comme suspectes aussi les naissances supérieures à 2 dans une année. Selon le cas on adoptera des méthodes de correction sans recours au fichier manuel ou au contraire avec retour au dossier.

La liste des incohérences inadmissibles doit être établie avec grand soin, et sera généralement assez brève. Celle des incohérences suspectes aura plutôt un caractère de test sur la validité de l'observation, et sera donc relativement brève elle aussi, ne portant que sur les caractères mal observés, et souvent davantage pour avoir une idée de la qualité de l'enquête que pour en rechercher une correction à tout prix.

1 -c.b.) Cohérence externe :

S'il existe des liaisons entre des diverses questions figurant dans un questionnaire, il existe aussi des liaisons d'une enquête à l'autre, ou d'un fichier à

... / ...

un autre, L'existence de données extérieures peut constituer un guide précieux pour l'examen de quelques questions pour lesquelles on ne dispose sans cela d'aucun critère objectif. Si, par exemple, lors d'une enquête antérieure on a pu constater que les salaires se hiérarchisent selon l'âge et qu'ils évoluent, à un âge donné, dans une certaine fourchette on pourra utiliser cette fourchette (éventuellement mise à jour) comme critère d'acceptabilité d'une réponse.

1 - d) Hiérarchie des contrôles :

Les contrôles ont un double but : d'une part éliminer du fichier toute impropreté formelle qui rendrait impossible la confection des tableaux, d'autre part éliminer les cas les plus invraisemblables qui risquent de fausser les calculs (moyennes écarts, etc...). Une enquête se déroulant en plusieurs étapes, terrain, chiffrement exploitation et analyse, lors de chaque étape devront avoir lieu des contrôles et chaque fois selon une hiérarchie différente. Sur le terrain et lors du chiffrement l'important est d'obtenir des renseignements les meilleurs possibles. L'intervention très rapide des moyens informatiques peut permettre une aide à ces contrôles, essentiellement parce qu'elle permet alors des retours sur le terrain. Par contre lors de l'exploitation proprement dite c'est essentiellement sur le plan formel que pourront et devront se dérouler les contrôles. Nous avons essayé de donner la liste des principaux contrôles à envisager. En pratique un certain nombre de choix devront être fait tant dans la quantité de contrôles que dans leur ordre. Cet ordre n'est en effet pas indifférent, d'autant plus que souvent l'on effectuera les corrections au fur et à mesure que seront détectées les erreurs. L'organisation qui nous semble la meilleure pour une grande enquête pourrait être la suivante :

- mise en ordre du fichier avec contrôle des identifiants, correction des identifiants erronés et remise en ordre définitive (surtout pour les fichiers hiérarchisés). (Le contrôle d'exhaustivité peut aisément se faire au cours de cette étape),
- contrôles logiques des principales variables, avec retour au fichier manuel pour les cas les plus aberrants, rejet dans la catégorie "non déclarée" pour les cas certainement erronés, mais de peu de conséquence (éventuellement correction automatique sur critère interne de ces cas),
- contrôle de validité de la totalité des codes et correction simultanée.

B - Corrections des erreurs.

Dans la partie consacrée aux contrôles nous avons défini deux types d'objectifs : vérifier qu'aucune erreur de forme ne subsiste dans le fichier afin d'en permettre l'exploitation et détecter dans la mesure du possible, les erreurs fondamentales, que celles-ci soient systématiques ou accidentelles. Ces opérations de contrôle sont indispensables, mais ne se suffisent pas en elles-mêmes : d'une

.../ ...

façon ou d'une autre les erreurs de forme devront être corrigées, et les erreurs de fond devront au moins être connues, même si elles restent difficiles à corriger.

Divers types de corrections peuvent être envisagés, soit directement par ordinateur (on parle alors de "correction automatique"), soit manuellement par substitution d'un enregistrement réputé exact à un enregistrement erroné. On utilisera l'un ou l'autre de ces types de corrections, souvent les deux, selon les contraintes propres à l'enquête : importance du fichier, possibilité du travail en atelier ou de retour sur le terrain, durée des opérations.

1 - Corrections automatiques.

Ce type de correction se fait par ordinateur, en principe sans intervention manuelle. Les erreurs rencontrées lors de la lecture d'un fichier pourront être systématiques (erreurs d'unités, décalages ..) ou aléatoires. Dans la mesure où elles pourront être détectées et ou l'on dispose des éléments permettant de les rectifier les erreurs systématiques peuvent être corrigées par procédés automatiques (changement d'unité par exemple), mais le plus souvent on préfèrera soit les traiter manuellement, soit les considérer comme aléatoires. Dans certains cas d'ailleurs une erreur systématique pourra subsister dans le fichier, aucune correction raisonnable n'étant possible : sous estimation évidente des salaires, oubli d'évènements trop anciens, etc...; lors de l'analyse des résultats il sera bien entendu nécessaire de connaître l'existence de ces biais, éventuellement d'en estimer l'importance.

Les erreurs aléatoires pourront être de tous les genres ; faute de frappe, omission d'un renseignement, code erroné, etc... Elles se trouveront dispersées dans le fichier sans ordre apparent. On fera donc généralement l'hypothèse que ces erreurs touchent des unités statistiques quelconques (il importe de s'en assurer) et les modes de corrections seront alors basés sur des principes de probabilité.

Finalement les procédés de corrections automatiques que nous allons exposer s'appliqueront donc aux erreurs aléatoires et éventuellement à certaines erreurs systématiques ou à des biais considérés comme aléatoires. En d'autre terme tout ce qui est considéré comme suspect et pour lequel on ne dispose pas d'informations valables (et dans la mesure où on ne recours pas à la correction manuelle) sera d'abord rejeté en "non déclaré" puis traité comme tel.

Il est à noter que le rejet en "non déclaré" de tout renseignement erroné ou suspect peut être considéré comme suffisant sur le plan formel, puisqu'on peut alors produire des tableaux à la seule condition que des cases "non déclaré" y figurent. Nous reviendrons en conclusion sur cette possibilité.

1 - a) Corrections alternatives.

Ce procédé est le plus simple, mais il n'est satisfaisant que quand des erreurs sont relativement rares. En cas d'erreurs on attribue alternativement chacune des valeurs possibles.

Exemple 1 : sexe non déclaré ; lorsque l'on rencontre un enregistrement présentant ce défaut on lui attribue le code 1 (masculin), puis à la seconde rencontre le code 2 (féminin), puis alternativement 1, 2, 1, 2, etc... On voit qu'on obtient une répartition par sexe en principe proche de la répartition réelle, les hommes et les femmes étant en nombre voisin dans la population, (il s'agit bien sûr d'une enquête sur la population totale).

Exemple 2 : état matrimonial non déclaré ; comme dans le cas précédent on attribue alternativement des valeurs 1 (célibataire), 2 (marié), 3 (veuf) et 4 (divorcé, séparé) puis à nouveau 1, 2, 3, 4, 1, 2, etc... Ce cas est déjà légèrement différent du 1er puisque la répartition dans l'ensemble de la population n'est pas aussi régulière. ON peut perfectionner le système en adoptant une pondération différente des corrections, par exemple 1, 2, 2, 3, 1, 2, 2, 4, etc..., soit 2 célibataires et 4 mariés pour 1 veuf et 1 divorcé, répartition déjà plus proche de la réalité. Cette pondération pourra de plus être différente selon l'âge.

L'inconvénient évident de ce procédé, par ailleurs de mise en oeuvre extrêmement simple, est que la fréquence réelle des états possibles n'étant pas connue à priori on doit en préjuger. Par ailleurs rien ne prouve que les erreurs soient effectivement aléatoires, donc que la population corrigée, artificiellement rendue semblable au reste de la population, l'était effectivement.

Pour l'application de ce procédé on prend généralement comme référence la répartition attendue dans l'ensemble de la population. Si une indication même subjective, permet de penser qu'en réalité les erreurs touchent une catégorie particulière de population on peut choisir une répartition volontairement biaisée. Si par exemple on a constaté chez les divorcés une tendance à ne pas déclarer leur état matrimonial on pourra adopter le cycle de correction 1, 4, 2, 4, 3, 4, etc... faisant apparaître 3 divorcés pour 1 célibataire, 1 marié et 1 veuf. Ce jeu a cependant ses limites, en particulier du fait que dans le cas envisagé la tendance sera plus fréquemment pour un divorcé de se déclarer marié que de ne pas répondre.

1 b) Correction par le contexte.

Pour certains caractères les liaisons entre plusieurs questions figurant dans l'enregistrement permettent de fixer de façon àpeu près certaine la réponse exacte. Par exemple à certains âges l'état matrimonial ne peut être que célibataire, ou le statut d'occupation qu'inactif (ou écolier). La correction est alors évidente.

Ce procédé ne peut toutefois qu'être assez partiel, car il est rare que plusieurs questions soient strictement redondantes ; même dans ce cas d'ailleurs se posera une question de priorité car s'il y a contradiction entre deux réponses laquelle doit être considérée comme exacte ?

Couplé avec la correctionalternative il a l'avantage d'éliminer d'éventuelles incohérences internes.

.../...

1 c) Correction par ratios.

Si les liaisons certaines entre caractères d'un même individu sont relativement rares, les liaisons entre un caractère et une quantité, ou entre 2 quantités sont beaucoup plus fréquentes. Ainsi par exemple la relation entre salaires et charges sociales (assurances, retraites, etc..) est assez rigide, celle entre âge, catégorie socio-professionnelle et salaire est plus souple mais réelle, etc... On peut alors éventuellement calculer l'élément manquant à partir de ceux dont on dispose ; connaissant le sexe, l'âge et la catégorie socio-professionnelle d'un individu on peut lui attribuer un salaire correspondant au salaire moyen des individus de mêmes caractéristiques.

Ce procédé est relativement difficile à mettre en oeuvre, par les calculs qu'il exige et par ses limites propres ; il est en effet nécessaire de disposer des "ratios" permettant les corrections. Parfois une source extérieure peut permettre de savoir à priori quels seront les ratios et il suffit alors de les appliquer ; mais dans le cas le plus général on devra les extraire de l'enquête elle-même : un premier passage du fichier permet de calculer par exemple le salaire moyen par C.S., sexe et âge, ppur les individus ayant déclaré ces 3 éléments et on applique lors d'un second passage ce salaire moyen aux individus ayant seulement déclaré CS et âge. On voit que s'il manque l'un de ces éléments on est arrêté, sauf à les "corriger" eux-mêmes.

1 d) Correction par "profils types"

Semblable dans son principe au cas précédent, mais portant aussi sur les caractères qualitatifs, on peut déterminer un état moyen pour un individu sur lequel on a assez peu d'informations : par exemple un homme de 45 ans sera "en moyenne" marié, salarié, etc... A partir de quelques données de base on pourra donc affecter à un individu des caractéristiques moyennes, qui auront surtout l'avantage d'être parfaitement banales.

L'établissement des ces profils types présente évidemment les mêmes inconvénients que précédemment : ou bien on les tire de renseignements extérieurs à l'enquête, mais alors correspondent-ils à la population étudiée, ou bien on les extrait de l'enquête elle-même, d'où nécessité de plusieurs passages.

1. e) Méthode du "HOT DECK"

Dans les quatre types de redressement automatique précédents la correction consistait à remplacer le renseignement erroné par un renseignement en moyenne exact, les méthodes 1a et 1b supposant que la distribution de référence soit connue (ou imposée), les méthodes 1c et 1d se référant à la distribution réellement observée. A l'exception de la première ces méthodes sont relativement difficiles à mettre en oeuvre , sans que cette difficulté soit compensée par une qualité incontestable.

.../...

Le procédé du "Hot Deck" allie les avantages d'une bonne méthode et d'une mise en oeuvre aisée.

Dans son principe le "hot deck" consiste, lorsque l'on rencontre un enregistrement erroné, à le remplacer par un autre enregistrement exact pris au hasard dans le fichier. Sur un fichier important les lois des probabilités pourront jouer, et l'enregistrement de remplacement correspondra en espérance mathématique au cas moyen. On aura donc par rapport au redressement alternatif supprimé les incohérences (tout enregistrement erroné est remplacé par un enregistrement existant au fichier et non erroné, donc cohérent) et l'arbitraire (la loi de remplacement n'est pas fixée par l'opérateur mais par le hasard).

En pratique, et cela constituera un avantage supplémentaire, on pourra procéder au redressement au fur et à mesure du déroulement du fichier : lorsque l'on rencontrera une erreur on prendra parmi les renseignements précédents l'enregistrement correctif.

On peut corriger soit la totalité de l'enregistrement, soit seulement une partie ou un seul élément. Le procédé reste sensiblement le même : à partir de 2 ou 3 caractères simples de référence (qui alors doivent tous être sans erreur) on sélectionne parmi les unités statistiques précédant celle à corriger celle qui en est le plus proche (par exemple même sexe, même âge et même état matrimonial) puis on attribue à l'enregistrement erroné le ou les caractères de l'unité ainsi sélectionnée.

L'inconvénient dans la mise en pratique est qu'il est alors nécessaire soit de revenir en arrière sur le fichier soit de conserver en mémoire quelques enregistrements parmi lesquels on procèdera à la sélection. Cet inconvénient sera supprimé si l'on se fixe les règles simples :

a) remplacement d'un enregistrement erroné par un enregistrement complet : la présence d'une seule erreur dans l'enregistrement justifie alors son rejet total ; l'enregistrement remplaçant pourra être celui le précédant immédiatement (il suffit donc de conserver toujours en mémoire un seul enregistrement, l'avant dernier lu). Si les erreurs sont réparties aléatoirement dans le fichier, les corrections sont aussi aléatoires, puisque déterminées par leur place.

b) correction d'un seul caractère erroné : si un caractère est erroné (par exemple le salaire) on recherche dans les enregistrements précédents le salaire d'un individu présentant par exemple même sexe, même groupe d'âge et même catégorie socio-professionnelle. Il faut alors constituer, et conserver en mémoire, un tableau "déformable" donnant pour chaque sexe, groupe d'âge et CS, le dernier salaire rencontré dans le fichier (tableau rempli arbitrairement avant le début du

... /...

déroulement du fichier, et qui se déformera de lui-même au fur et à mesure de la lecture). Lorsque l'on rencontre un salaire erroné on lit l'âge, le sexe et la CS de l'individu concerné, on recherche dans le tableau un individu de même sexe, âge et CS et on remplace le salaire erroné par celui lu dans le tableau.

Si l'on adopte cette correction d'un seul caractère, il faudra faire plusieurs opérations, une pour chacun des caractères susceptibles d'être corrigés, ce qui peut entraîner la construction d'un nombre assez considérable de tableaux de référence, donc un encombrement non négligeable de la mémoire.

Le rapporchement de la méthode du hot deck avec la méthode des sondages est évident : les corrections sont prises au hasard dans l'ensemble des réponses exactes. Pour respecter strictement ce hasard on aurait intérêt à travailler sur des fichiers les plus désordonnés possibles ; en réalité un "bon" ordre du fichier est souhaitable, correspondant à l'idée de stratification : si le fichier est rangé dans un ordre géographique, par exemple, l'individu sélectionné pour corriger une erreur présentera, outre les caractères communs recherchés (sexe, âge, etc...) une proximité géographique.

2 Redressements manuels

Malgré leur limite les redressements automatiques présentent de très grands avantages pratiques et doivent être utilisés chaque fois que cela sera possible. Leur inconvénient majeur tient à leur absence de souplesse, inhérent à leur définition. Par ailleurs, ils doivent en principe être réalisés à un moment bien précis de l'exploitation, après l'introduction du fichier en machine et avant la production des tableaux.

Beaucoup plus souples, et pouvant être réalisées lors de toutes les étapes de l'exploitation (y compris, bien que cela nous paraisse condamnable, après la production des tableaux), les corrections manuelles seront généralement plus difficiles à mettre en oeuvre.

Tenant compte des avantages certains de la rigidité (la correction ne dépend pas de l'opérateur) et de la rapidité de la correction automatique et des avantages de souplesse de la correction manuelle un certain nombre de choix seront à faire pour établir le plan de redressement. Généralement on procèdera à une correction automatique pour les questions d'importance relativement secondaire et les erreurs plus fondamentales, soit en raison de l'importance de la question soit à cause du poids du questionnaire seront traitées à la main.

.../...

Dans une enquête sur la gestion financière des entreprises par exemple on pourra décider de faire deux lots de questionnaires, l'un concernant les grandes entreprises, peu nombreuses mais à chiffre d'affaires élevé, qui seront éventuellement corrigées à la main, l'autre concernant les petites entreprises, beaucoup plus nombreuses, pour lesquelles on corrigera automatiquement la plupart des erreurs et manuellement certaines erreurs plus conséquentes (masse des salaires par exemple).

2 . a. Redressement à la saisie

Ce procédé de redressement est lié à la nature du matériel informatique dont on peut disposer. Il est en effet nécessaire de pouvoir vérifier la validité d'une information au moment même où elle est saisie (transferée du document manuscrit de base sur un support, carte ou bande magnétique, exploitable par l'ordinateur), et éventuellement de signaler la non-validité à l'opérateur afin qu'il la corrige. Un jeu de claviers de perforation associés à des écrans permettant la lecture de messages ou a une "imprimante" le tout lié à un ordinateur est donc nécessaire.

Lorsque l'on dispose d'un tel matériel beaucoup de contrôles sont possibles : acceptabilité de l'identifiant (si celui-ci a une clef), non dépassement de code, ratios convenables etc... Quand une valeur de code, ou une quantité, vient d'être saisie l'ordinateur peut en effet vérifier que la valeur est acceptable, que la quantité entre dans une "fourchette" préétablie etc... et si tel n'est pas le cas le signaler à l'opérateur. Celui-ci disposant en principe du dossier pourra selon le cas retranscrire le code exact s'il s'agissait d'une erreur de frappe, le rechercher s'il s'agissait d'une erreur de chiffrement etc.. On voit cependant tout de suite la lourdeur de l'opération car s'il ne s'agit pas d'une faute de frappe l'opérateur doit décider lui-même de codes ou valeurs de remplacement, ce pour quoi il n'est pas obligatoirement compétent. De plus le programme de recherche des erreurs, lourd, immobilisera une partie importante de l'ordinateur pendant toute la durée de la saisie, qui peut s'étendre sur plusieurs mois.

En pratique on n'utilise le plus souvent ce procédé que pour des contrôles relativement simples et pour lesquels la correction éventuelle est élémentaire, ou peut être différée afin de ne pas immobiliser la chaîne de saisie trop longtemps (contrôles d'exhaustivité, d'identifiants, de structure, codes simples).

.../...

2. B. Redressement en ligne

Nécessitant le même matériel que précédemment ces types de redressement sont plus riches, mais demandent une participation plus poussée de l'ordinateur. La manipulation initiale étant la même l'opérateur introduit une valeur de code, l'ordinateur vérifie la validité de ce code (non dépassement ou cohérence avec d'autres éléments du questionnaire) et en cas d'erreur proposera une ou plusieurs solutions (par redressement automatique).

L'opérateur alors pourra choisir l'une des solutions proposées ou éventuellement en proposer une lui-même.

Souvent on profite de ces opérations pour consulter automatiquement des nomenclatures : l'opérateur frappera par exemple en clair l'activité économique et l'ordinateur donnera lui-même le numéro correspondant, s'il existe, ou demandera des précisions s'il y a ambiguité.

On voit que pour utiliser les deux procédés ci-dessus il sera nécessaire d'une part de consentir à l'immobilisation d'un matériel très important, d'autre part, de disposer d'un corps d'opérateurs au courant de l'enquête et capable d'initiative. Afin de pallier les difficultés qui pourraient naître il sera toujours nécessaire de prévoir la possibilité d'un rejet provisoire du questionnaire erroné, ce rejet permettant aux opérations de continuer par ailleurs.

2. c. Redressement différé

Plutôt que de chercher à redresser le fichier au moment de la saisie, gain de temps qui se traduira souvent par un investissement très lourd en personnelet en matériel, et qui ne se justifie pas toujours par son efficacité, on préfère souvent produire des listes d'erreurs que l'on corrigera à tête reposée.

La procédure est alors la suivante : lorsqu'une erreur est détectée, l'enregistrement correspondant est soit exclu du fichier soit placé en réserve, un message d'erreur est émis, à l'aide de ce message on recherche le dossier correspondant, on le corrige et l'on remplace alors l'ancien enregistrement erroné par un nouvel enregistrement corrigé. Il est souvent plus interessant d'annuler la totalité de l'enregistrement erroné (partie exacte et partie erronée que de n'en corriger qu'une partie. Le risque est toutefois alors d'introduire une nouvelle erreur dans la partie saine.

Le redressement différé est incontestablement la meilleure formule pour corriger un fichier, puisqu'il permet l'examen cas par cas des erreurs, et leur correction (y compris par retour sur le terrain). Il présente cependant un premier inconvénient par sa formule même : on doit retourner, parfois assez longtemps après, à des dossiers déjà exploités ; souvent alors le personnel de l'enquête a été dispersé, parfois certains dossiers ont été égarés, les renseignements sont trop anciens pour être vérifiés sur le terrain etc...

.../...

D'autre part il faudra créer un nouveau fichier correctif,qui lui aussi présente des risques d'erreurs,. Enfin la procédure est très longue, pour un bénéfice, souvent illusoire.

2. d. Messages d'erreurs

En tout état de cause les erreurs détectées et corrigées doivent être décomptées. D'autre part si l'on procède à un redressement différé il sera nécessaire de travailler sur des listes d'erreurs afin de pouvoir les corriger. On devra donc dans tous les cas émettre des messages faisant ressortir les anomalies rencontrées. Ces messages doivent permettre à la fois le décompte et la correction des erreurs. Ils devront donc :

- permettre d'identifier sans difficulté les enregistrements erronés, d'abord pour retrouver les dossiers correspondants, ensuite pour revenir à l'enregistrement mis en cause ;

- permettre d'identifier l'erreur trouvée, en la signalant très clairement (ex : manque salaire, charges sociales trop élevées, etc...) ;

- reproduire la donnée suspecte, celle-ci pouvant servir à la correction (si par exemple il s'agit d'une erreur de perforation la seule indication "salaire trop élevé" est incompréhensible sur le dossier).

D'autres qualités sont demandées aux messages d'erreurs, essentiellement d'ordre pratique : maniabilité, classement selon le type d'erreur, homogénéité des indications fournies etc...

Dans le cas de redressement à la saisie, ou en ligne, ces conditions sont automatiquement remplies puisque c'est au moment où l'on rencontre l'erreur qu'on la signale et la corrige. Les messages d'erreurs n'auront alors qu'un intérêt d'archivage et peuvent à la rigueur se limiter au décompte de ces erreurs. Dans le cas du redressement différé peut se poser un problème pratique gênant : doit-on signaler toutes les erreurs d'un enregistrement ou tous les enregistrements correspondant à un type d'erreur. Dans le premier cas le dossier ne sera sorti qu'une seule fois et corrigé entièrement mais l'organisation de l'atelier de correction peut s'en ressentir puisque d'un document à l'autre des erreurs de type très différent seront rencontrées. Dans le second cas on pourra au contraire corriger, type d'erreur par type d'erreur, mais les manipulations seront plus nombreuses.

.../...

3 Conclusions sur les redressements d'erreurs.

Un fichier présentera toujours des erreurs de forme et de fond, et les moyens informatiques offrent la facilité de détecter un certain nombre de ces erreurs.

Pour ce qui est des erreurs formelles il faudra toujours les corriger, d'une façon ou d'une autre, car leur maintien entraînerait l'impossibilité de fournir les tableaux, objectif même d'une exploitation. Les autres erreurs, souvent plus fondamentales, posent à la fois un problème théorique et un problème pratique : la mise en oeuvre d'un système de détection et de correction est fort onéreuse, et peut-on, ou même doit-on, remplacer une réponse suspecte par une autre plus satisfaisante apparemment, mais arbitraire ? Le purisme consisterait à ne procéder qu'à des corrections neutres, formellement acceptables, qu'on peut en gros ramener à l'utilisation du système de cases "non déclaré" dans les tableaux. Cette attitude nous paraît à proscrire, d'abord parce qu'elle entretient l'illusion que tout ce qui est déclaré est exact, ensuite parce qu'elle n'a aucun intérêt pratique : l'habitude est prise depuis fort longtemps de lire la partie saine des tableaux et de négliger ou de répartir proportionnellement le contenu des cases "non déclarées". Un mode de correction relativement simple donne les mêmes résultats, sans l'inconvénient de "traîner" en permanence des tableaux incomplets. Il reste que la perfection formelle des tableaux corrigés est trompeuse. Il importera de ne pas oublier que cette apparente perfection n'a pu être obtenue que par l'élimination d'un certain nombre d'erreurs.

Contrôles des résultats :

Le fichier ayant été rendu propre sur le plan formel et débarrassé dans la mesure du possible des erreurs individuelles flagrantes, il reste à produire des tableaux. Généralement on ne lancera pas tout de suite un programme très lourd des tableaux mais on commencera par l'édition de quelques tableaux de contrôle :

1 - Tableaux de décompte d'erreurs

Lors de la mise au propre du fichier on a détecté un certain nombre d'erreurs que l'on aura corrigé. Les erreurs matérielles (erreur de perforation par exemple, erreur d'unité, etc...) n'ont plus grande importance dès lors qu'on a pu les rectifier. Il est cependant utile d'en connaître le nombre car il est certain que le nombre d'erreurs de ce type non détectées est proportionnel à celui des erreurs détectées. Si par exemple on a trouvé 10 % d'individu au sexe différent de 1 (masculin) ou 2 (féminin) ces erreurs ne peuvent guère provenir que d'une mauvaise perforation et il est alors probable que pour les codes

plus complexes le nombre d'erreurs de perforation sera important. Les erreurs de fond par contre (par exemple absence systèmatique de réponse à certaines questio questions, confusion entre nombre d'enfants nés vivants et nombre d'enfants actuellement en vie etc...) doivent faire l'objet de tableaux détaillés qui permettront à l'analyste de proposer des explications, ou qui dans tous les cas devraient lui permettre d'éviter des interprétations erronées car les corrections apportées au fichier dans ces cas sont toujours plus ou moins arbitraires, et le biais introduit par ces corrections sera négligeable si elles sont rares, considérable si elles sont nombreuses.

On devrait toujours constituer pour chaque enquête un dossier comportant un tableau statistique de décompte des erreurs détectées, le mode de correction adopté et éventuellement une liste des questionnaires comportant de graves erreurs,. En pratique de tels dossiers sont assez rares et toujours confidentiels !

2 - Tableaux des données brutes

Dans la mesure où le fichier constitué le permet, avant tout contrôle et toute correction, il est souvent judicieux de produire quelques tableaux assez simples (répartition par sexe et âge, distribution des salaires selon le sexe par tranches, nombre d'enfants selon l'âge des mères etc...) Ces tableaux présenteront l'avantage de permettre rapidement de détecter des anomalies considérables (par exemple confusion des salaires mensuels et annuels) et de faire apparaître dans les cases prévues à cet effet (non déclaré, autres cas...) l'importance des erreurs probables. Ils permettront aussi, par comparaison avec des tableaux définitifs de mesurer l'influence de corrections que souvent on ne maitrise pas complètement. Eventuellement on peut produire ces tableaux à partir d'un échantillon relativement restreint de questionnaires.

3 - Tableaux de contrôles

A partir du fichier propre la production de tableaux extrêmement simples donnera les principaux résultats. Une analyse rapide de ces résultats permettra d'une part d'estimer, intuitivement peut-être, la vraissemblance des résultats d'ensemble, d'autre part de décider rapidement du degré de finesse que l'on pourra rechercher dans les résultats détaillées.

F. PRADEL de LAMAZE

Chapitre III : Codification

A partir d'une population donnée l'objet de la statistique est de classer les individus formant cette population selon leurs caractères. Ces caractères pourront se présenter de façon claire, quasi évidente, et avec un nombre de modalités restreintes (ou en tout cas dénombrables) comme dans le cas du sexe, de l'âge, du nombre d'enfants etc..., ou de façon beaucoup plus complexe. Dans l'un comme dans l'autre cas le statisticien devra établir une nomenclature, liste exhaustive et ordonnée des modalités possibles, et pour les besoins de l'exploitation, particulièrement du point de vue informatique, à cette nomenclature devra être associé un code, tel qu'à chaque cas envisagé dans la nomenclature corresponde un signe (généralement un nombre) et un seul.

Etablir une nomenclature est généralement une opération délicate, dès lors qu'il ne s'agit pas de caractères simples. L'objet de ce manuel n'étant pas la théorie statistique nous ne développerons pas ce point, mais il est cependant nécessaire de rappeler quelques contraintes :

a) une nomenclature n'a de raison d'être que si elle est opératoire. L'exhaustivité des cas possibles est donc nécessaire, mais le problème est le plus souvent de contracter d'une façon ou d'une autre le nombre de ces cas. Enumérer par exemple tous les emplois possibles et affecter à chacun un numéro d'ordre de 1 à n, n pouvant alors être de l'ordre de centaines de milliers, n'a aucune raison d'être puisque sur un tableau statistique une centaine d'éventualités semble un maximum. Dans ce cas donc il s'agira de regrouper, par proximité, tous les emplois se "ressemblant". Malheureusement la ressemblance sera rarement évidente et pourra varier selon le point de vue d'où l'on se place : regrouper d'une part les emplois de bureau, d'autre part les emplois ouvriers etc... peut être très légitime pour une étude selon la nature de l'emploi occupé, mais perd beaucoup de son intérêt si l'on envisage par exemple l'étude de la dispersion des salaires.

b) une nomenclature doit avoir une certaine permanence dans le temps et l'espace. Une étude isolée est certes interéssante mais le plus souvent ne prend sa pleine dimension que dans la comparaison, soit avec une situation passée soit avec une autre population.

.../...

D'une date à l'autre ou d'un pays à l'autre beaucoup de choses évoluent ; pour reprendre l'exemple des emplois certains de ceux-ci disparaissent avec le temps, d'autres apparaissent, et une tendance à la spécialisation plus ou moins poussée se fait jour un peu partout. Conserver longuement une nomenclature, aussi bonne soit-elle, sera donc généralement difficile ; la changer à toute occasion sera extrêmement dangereux .

c) Une nomenclature doit tenir compte à la fois de l'instabilité de certains caractères et de la précision de l'enquête. Il serait ainsi illusoire de prétendre observer de façon très détaillée certaines caractéristiques si dans la population ces caractéristiques ne représentent pas une situation bien définie, ou si la nature même de l'enquête ne permettait pas de distinction très fine. L'emploi pourra ici aussi servir d'exemple : dans la plupart des pays en voie de développement surtout de très nombreux métiers plus ou moins marginaux sont occupés pendant des périodes très brèves par certains individus ; recenser tous ces emplois marginaux serait assez vain (sauf au cas où l'enquête aurait justement cette fin), les classer et prétendre ainsi représenter l'état de la population serait très probablement erroné.

Afin de tenir compte de ces contraintes on s'efforcera donc d'établir des nomenclatures relativement détaillées au départ, permettant divers regroupements. Des méthodes plus ou moins astucieuses, dont l'essence apparait dans les nomenclatures "emboitées", ont été utilisées à l'époque de la mécanographie. Actuellement on fait plus souvent appel à des "tables de passage" qui permettent, à partir d'une nomenclature fine de reconstituer rapidement telle ou telle nomenclature agrégée. En pratique pour réaliser des exploitations informatiques on devra toujours prévoir, de façon automatique, une étape de "recodification".

La procédure est alors la suivante : à partir d'un questionnaire de base on procède au chiffrement (opération consistant à partir de l'information littérale à la transformer en un nombre ou éventuellement en signe alphabétique) détaillé de chacun des caractères. Le détail de ce chiffrement est alors fonction essentiellement de la précision de l'enquête, sans préjuger ni de la forme des tableaux ni de leur nombre de lignes ou de colonnes ; par contre il devra tenir compte de certaines contraintes concernant les modes de regroupement possibles : par exemple si l'on chiffre la branche d'activité à laquelle appartient une entreprise on devra pouvoir isoler les coopératives agricoles qui dans certains regroupements pourront être agrégés à l'agriculture, dans d'autres à l'industrie ou même au commerce. A partir de ce chiffrement, après mise sur bande magnétique, contrôle et éventuellement correction, on réaffecte à l'individu l'ensemble des codes correspondant à chacune des nomenclatures définitives envisagées, et ceci en faisant appel soit à un seul caractère soit parfois à plusieurs. Les exemples suivants illustrent cette procédure :

.../...

Exemple 1 : Dans les cas les plus simples (caractères qualitatifs non ambigus), nomenclatures et codes sont établis très rapidement : masculin = 1, féminin = 2 ou célibataire = 1, marié = 2, veuf = 3, divorcés séparés = 4, etc.

Exemple 2 : Un cas relativement plus délicat est celui où les individus peuvent appartenir à plusieurs classes. Dans ce cas, il est généralement préférable d'établir une hiérarchie des classes et de ne placer l'individu que dans la classe la plus haute. On peut aussi utiliser un système de codification un peu complexe mais qui n'entraine pas de perte d'information (code binaire).

Exemple : Diplômes d'instruction générale

Système 1 :

Néant	0
Sait lire et écrire	1
Certificat d'étude	2
BEPC	3
Bacc. ou plus	4

Chaque individu étant classé selon son niveau le plus haut.

Système 2 :

Néant	0
Sait lire et écrire	1
Certificat d'étude	2
BEPC	4
Bac. et plus	8

Chaque individu est codé selon la somme des codes correspondants aux diplômes dont il dispose. Ainsi tout individu codé 1, 3, 5 ou 9 sait lire et écrire ; tout individu codé 5 sait lire et écrire et n'a que le BEPC ; tout individu codé 7 sait lire et écrire et a le certificat d'études et le BEPC.

La décomposition de tout nombre en puissances successives de 2 étant unique on peut grâce à ce système décompter aussi bien les individus ayant un diplôme donné quels que soient leurs autres diplômes, que ceux combinant plusieurs diplômes.

Exemple 3 : Dans le cas le plus général, le nombre de possibilités est très grand. Il faut alors constituer des nomenclatures "emboitées", telles que l'on puisse procéder à des regroupements successifs, du caractère le plus détaillé au caractère le plus général. Un système de codification du type décimal s'adapte généralement assez bien à ce type de nomenclature : dans un tel système, le 1er chiffre rassemble tous les individus appartenant à un grand groupe, les deux premiers permettent de distinguer les principaux sous-groupes, 1 ou 2 chiffres supplémentaires permettent d'atteindre le détail le plus fin.

.../...

Voici par exemple la nomenclature des professions (extrait) utilisée pour le recensement de l'Algérie (1966)

GROUPES	PROFESSIONS	CODES
GROUPE 0	PERSONNES EXERCANT UNE PROFESSION LIBERALE ,TECHNICIENS ET ASSIMILES.	
	Architectes, ingénieurs et géomètres	00
	Chimistes, physiciens, géologues et autres spécialistes des sciences physiques	01
	Biologistes, vétérinaires, agronomes et spécialistes exerçant des professions connexes	02
	Médecins, chirurgiens et dentistes	03
	Infirmiers et sage-femmes	04
	Spécialistes et techniciens paramédicaux	05
	Personnel enseignant	06
	Prêtres et membres assimilés d'ordre religieux	07
	Juristes	08
	Artistes, écrivains et assimilés	09
	Dessinateurs et techniciens des sciences physiques et des sciences appliquées	0X
	Autres personnes exerçant une profession libérale, techniciens et assimilés	0Y
GROUPE 1	DIRECTEURS ET CADRES ADMINISTRATIFS SUPERIEURS	
	Directeurs et cadres supérieurs de l'Administration publique	10
	Directeurs, cadres administratifs supérieurs et propriétaires exploitants	11
GROUPE 2	EMPLOYES DE BUREAU	
	Aides comptables, teneurs de livres et caissiers	20
	Sténographes et dactylographes	21
	Autres employés de bureau	29
GROUPE 3	VENDEURS	
	Propriétaires exploitants (commerce de gros et détail)	30
	Agents d'assurances, agents immobiliers, démarcheurs de banque, agents de vente de service, et vendeurs aux enchères, courtiers maritimes, apprêteurs	31
	Voyageurs de commerce, représentants et placiers	32
	Commis, vendeurs, employés et travailleurs assimilés	33
GROUPE 4	AGRICULTEURS, PECHEURS, CHASSEURS, FORESTIERS ET TRAVAILLEURS ASSIMILES	
	Agriculteurs et directeurs d'exploitations agricoles	40/41
	Travailleurs agricoles	42
	Chasseurs et travailleurs assimilés	44
	Pêcheurs et travailleurs assimilés	45
	Bûcherons et autres travailleurs forestiers	46
GROUPE 5	MINEURS, CARRIERS ET TRAVAILLEURS ASSIMILES	
	Mineurs et carriers	50
	Foreurs de puits et travailleurs assimilés	51
	Ouvriers spécialisés dans l'enrichissement des minerais	52

.../...

**

Exemple : Nomenclature des catégories socio-professionnelles (extrait)

Branche d'activité	Statut	Profession	Dimension de l'entreprise	Catégorie professionnelle (CSP)	Code CSP
Agriculture	Propriétaire	Agriculteur	→	Agriculteur Exploitant	10
	Aide familial	"			
	Fermier	"			
	Métayer	"			
	Salarié	"		Sal. agricole	11
Industrie	Indépendant	Directeur	6 sal et +	Industriel	20
		"	0 à 5 sal.	Artisan	21
	Aide familial		6 sal et +	Industriel	30
			0 à 5 sal.	Artisan	31
	Salarié	Directeur		Cadre supér.	40
		Contremaitre	→	Cadre moyen	50
		Ouvrier		Ouvrier	60

Le caractère n'est pas pris en compte pour la détermination de la CSP.

Exemple 5 :Sur le questionnaire initial et sur la carte perforée correspondante, on a codé l'âge exact selon un code à 2 chiffres (25 = 25 ans révolus). On sait que certains tableaux devront être produits pour des regroupements d'âges quiquennaux, ou décennaux, d'autres seulement pour certains âges, etc... Lors du transfert de la carte perforée sur bande magnétique, on va donc transformer ce code "âge" en divers codes "âges regroupés", qui serviront de critères ligne (ou colonne)" pour l'établissement des tableaux, selon le système suivant :

** Exemple 4 :: Dans certains cas on a à synthétiser dans un même code deux ou plusieurs caractères que peuvent présenter les individus. Ce genre de problème a intérêt à être traité de façon rigoureusement systématique à l'ordinateur, mais peut l'être éventuellement à la main. Le traitement par ordinateur suppose bien entendu qu'au préalable ait été chiffré chacun des caractères composants :

... / ...

Age détaillé	AR1	AR2	AR3	AR4
...				...
11		01	01	11
12				12
13	...			13
14				14
15			02	15
16	04	02	03	16
17			04	17
18			05	18
19			06	19
20			07	
21	05	03	08	20
22			09	
23			10	
24			11	
25				
26	06	04	12	21
27				
28				
29				

Age détaillé	AR1	AR2	AR3	AR4
30				-
...	...	...	...	
61				
62				
63				
64				
65				
66	14			
67				
68				
69				
70		12	20	
71	15			
72				
73				
74				
75				
76				
-	-			

Chapitre IV

Organisation des ateliers

Il comprendra trois parties :

- le rôle de l'atelier manuel

- la place de l'atelier manuel dans la chaîne d'exploitation

- l'organisation optimale de l'atelier

4 - 1 - Le rôle de l'atelier manuel

Ne pas oublier que le rôle de l'atelier manuel ne se limite pas au seul chiffrement des documents d'enquête (c'est d'ailleurs pourquoi le nom d'"atelier manuel" paraît plus adapté que celui d'"atelier de chiffrement"). L'atelier manuel intervient dans les phases suivantes de l'exploitation :

- préparation de l'enquête : . tirage de l'échantillon : le tirage lui-même peut être fait de façon plus ou moins automatique (on donnera rapidement quelques exemples comme le tirage des échantillons d'enquêtes-ménage ou la gestion du FILE) ; l'intervention de l'atelier manuel consiste à :

. effectuer le tirage proprement dit de l'échantillon (choix des unités statistiques enquêtées)

. rédiger les fiches-adresses à remettre aux enquêteurs (localisation et caractéristiques connues de l'unité à enquêter)

. mettre sous enveloppes (choix des questionnaires en fonction de la nature de l'unité à enquêter) et expédier les documents en cas de collecte effectuée par voie postale

- gestion de l'enquête et premiers traitements : vérification de l'exhaustivité de la collecte (pointage des documents de retour, expédition de lettre de rappels, demandes de mise au contentieux) ; cette partie du travail peut être, dans une certaine mesure, automatisée (fichier de gestion d'enquête mis à jour au fur et à mesure de l'arrivée des questionnaires, d'où on peut tirer des statistiques sur la rentrée des documents, l'édition automatique des lettres de rappel, etc ...)

. contrôles et vérifications sommaires (comptages, contrôle systématiqu des variables-pivots) déclenchant éventuellement un retour à l'enquêteur ou une demande de renseignements complémentaires à l'unité enquêtée : cette phase sera indispensable si l'organisation de la chaîne de production suppose une introduction des données sans "chiffrement" préalable (exemple : certaines enquêtes auprès d'entreprises).

- chiffrement proprement dit : . traduction en codes des réponses données en clair par les unités statistiques enquêtées, à l'aide de codes, de nomenclatures, voire de fichiers administratifs (exemple de l'utilisation de SIRENE par le chiffrement de l'activité économique et du lieu de travail dans le RP) : on évoquera ici certains problèmes liés à la constitution de codes et de nomenclatures adaptés au travail de chiffrement (frontière avec le chapitre 3 consacré à la codification) ; on évoquera aussi la possibilité de consulter ces codes et nomenclatures à partir de terminaux utilisés en mode conversationnel

. contrôles de validité et de cohérence des réponses faites ; erreurs et anomalies : correction des erreurs détectées (voir § 4-2 ci-après) ;

. vérification du travail de chiffrement : de la nécessité d'une telle vérification dans certains cas, possibilité d'utiliser les méthodes statistiques de contrôle de fabrication (mesure de la qualité par prélèvement d'un échantillon et jugement sur échantillon)

. gestion du travail : vérification de l'exhaustivité du travail de chiffrement

- dépouillement : l'atelier manuel peut être amené à effectuer lui-même le dépouillement 'à la main' de l'enquête dans certains cas (résultats préliminaires : exemple de la population "légale" du RP ; comptages, tableaux simples), soit parce que les comptages constituent en eux-mêmes le résultat recherché, soit pour avoir des résultats très simples, mais rapides, soit pour effectuer certaines vérifications d'hypothèses nécessaires à la poursuite du traitement.

4 - 2 - La place de l'atelier manuel dans la chaîne d'exploitation

Plusieurs solutions sont possibles pour l'organisation de l'exécution des contrôles et redressements informatiques. La place de l'atelier manuel dans la chaîne d'exploitation dépend du mode d'exécution des contrôles/redressements qui a été retenu :

a) contrôles/redressements automatiques : lorsque une erreur (incohérences entre codes, codes non valides, absence de réponses, etc ...) est détectée, la correction de l'erreur est entièrement automatique, sans retour au document de base, ni a fortiori à l'unité enquêtée ; elle se fait à partir du reste de l'information reconnue valide pour l'unité statistique enquêtée, ou à partir des informations de même nature d'unités statistiques ressemblantes (méthode du hot-deck) : cf. chapitre 2 consacré aux contrôles et aux redressements

b) contrôles/redressements itératifs :lorsqu'une erreur, ou même une simple anomalie (qui ne sera pas forcément une véritable erreur) sera détectée par le programme informatique, un message d'anomalie est édité et expédié à l'atelier manuel ; celui-ci, après retour au document de base, et, le cas échéant, retour auprès de l'unité statistique enquêtée (par téléphone, par lettre, voire par envoi d'un enquêteur), doit proposer une mise à jour du fichier des données (correction des zones incriminées) ou confirmer que la réponse proposée est bonne et doit être conservée. Au bout d'un certain nombre d'itérations, le fichier est réputé correct (bon pour l'exploitation des données).

c) contrôles/redressements interactifs : les agents de l'atelier manuel disposent d'un outil de dialogue avec l'ordinateur : (1) terminal, dit "conversationnel", muni d'un clavier d'entrée des données et d'un écran sur lequel sont visualisées, d'une part les données rentrées, et, d'autre part le diagnostic effectué à l'aide d'un programme de contrôle des données implanté sur l'ordinateur (données non valides, données incohérentes entre elles ou ne respectant pas certaines contraintes a priori, données susceptent, ...).

Une variante à l'utilisation de terminaux conversationnels reliés à un ordinateur consiste à utiliser des mini-ordinateurs de gestion spécialement programmés à cet effet. Il n'est pas possible de proposer de réponse générale à la question de savoir s'il vaut mieux utiliser l'un ou l'autre des deux procédés conversationnels (terminaux ou mini-ordinateurs) : une étude coûts-avantages est à réaliser cas par cas.

./.....

---

(1) Bien qu'en pratique, l'étape de contrôle des données soit logiquement distincte de l'étape de chiffrement/codification, ces deux étapes sont, lorsqu'un terminal conversationnel est utilisé, effectuées simultanément (cf. COLIBRI).

La méthode de contrôles/redressements automatiques est à préconiser lorsque les unités statistiques interrogées sont suffisamment nombreuses et de 'poids' analogue (c'est très généralement le cas des enquêtes auprès des ménages). Une erreur de codification ou de saisie - qui sera redressée de telle sorte que la distribution statistique des unités sans erreurs ne soit pas modifiée - n'aura pas alors de répercussionsimportante. Il convient cependant de maîtriser le taux d'erreur commis par les agents chargés de la codification et de la saisie (des taux de l'ordre de 1 pour mille à 1 % paraissant admissibles).

En revanche, dès que les unités statistiques enquêtées sont de 'poids' trop différent (par exemple, entreprises dont la taille va de 0 salarié jusqu'à plusieurs milliers), les méthodes automatiques doivent être évitées. Le choix entre les méthodes itératives ou interactives dépendent alors de plusieurs facteurs, dont le principal est la possibilité pour l'agent chargé du contrôle de répondre en 'temps réel' ou pas à la détection de l'anomalie : si la correction d'une erreur nécessite le retour auprès de l'unité statistique interrogée, il vaut mieux éviter le recours aux méthodes interactives.

Ne pas oublier que :

- il est toujours possible de combiner deux ou trois de ces méthodes (par exemple : utilisation de méthodes automatiques pour les entreprises de moins de x salariés et de méthodes itératives pour les autres ; détection des anomalies en 'batch' méthode itérative - mais 'entrée' des corrections par méthode interactive de manière à s'assurer de la validité formelle des corrections faites)

- les méthodes interactives peuvent, en outre, être utilisées pour la consultation des codes et nomenclatures.

4 - 3 - L'organisation de l'atelier

L'organisation proprement dite de l'atelier est fonction de la réponse apportée à un certain nombre de questions :

a) faut-il centraliser le travail manuel ou le décentraliser sur une base géographique ?

La réponse à cette question est fonction :

- de la taille du pays (elle ne se pose que si la taille du pays est supérieure à quelques millions d'habitants - 4 ou 5 pour fixer les idées)

- de son infrastructure administrative et statistique (existence ou non de bureaux régionaux)

- de la taille de l'enquête (l'exécution du chiffrement d'un recensement peut nécessiter l'ouverture d'ateliers régionaux même si l'office statistique est entièrement centralisé)

- du mode de collecte (le chiffrement sera plus facile à centraliser si la collecte est effectuée par voie postale que si elle est faite par enquêteurs ; dans ce dernier cas, la proximité de l'enquêteur - pour la correction des erreurs détectées - est un facteur positif)

- de la complexité des consignes de chiffrement (si ces consignes sont complexes, les risques d'hétérogénéité en cas d'existence de plusieurs ateliers sont importants ; dans ce dernier cas, si on souhaite malgré tout décentraliser le travail - par exemple pour ne pas créer des unités de taille trop importante -, il faut songer à une décentralisation sur une base autre que géographique, par exemple, sur une base sectorielle pour les enquêtes auprès des entreprises)

b) <u>faut-il spécialiser les agents</u> ?

On peut songer à spécialiser les agents par type d'enquête, c'est-à-dire essayer de regrouper dans une même unité les agents chargés d'enquêtes analogues (par exemple enquêtes ménages, enquêtes entreprises, observation des prix de détail, ces trois domaines pouvant être aux-mêmes fractionnés ; par exemple, 'enquête ménages' découpées en 'démographie-emploi' et 'conditions de vie').

Une telle spécialisation peut poser des problèmes de plan de charge.

c) <u>faut-il découper la chaîne de traitement</u> ?

c'est-à-dire créer plusieurs équipes qui interviendront successivement tout au long du traitement d'un questionnaire :

. réception - pointage - premiers traitements

. un ou plusieurs postes de chiffrements (selon la structure du questionnaire

. vérification du travail

auquel on ajoutera une équipe 'horizontale' de gestion et de suivi du travail.

La méthode du poste de travail unique présente des avantages :

. meilleure motivation des agents qui ont une vue d'ensemble du travail à exécuter

. absence de ruptures dans la chaîne

. plus grande souplesse

et des inconvénients :

. nécessité d'avoir des agents de niveaux de formation homogènes (risque d'une dégradation de la qualité si ce n'est pas le cas)

. nécessité de multiplier les documents (codes, nomenclatures) nécessair au travail

. difficulté pour faire vérifier le travail accompli par les mêmes agents que ceux chargés de son exécution.

Il n'y a, en fait, pas de réponse générale à la question posée, la réponse dépendant avant tout du niveau de formation des agents qui composent l'atelier

d) <u>ne pas négliger les problèmes de planification et d'ordonnancement des travaux</u>

- évaluer au mieux les temps nécessaires (utilisation de temps unitaires types ; exécution d'un dépouillement pilote ; ne pas oublier les 'frais généraux' : manipulation, transmission d'un poste à un autre, classement-archivage, etc ...)

- bien organiser les transmissions de documents avec l'aval (saisie et dépouillement informatique) ; bien définir les lots de travail et leur rythme de traitement ; précautions particulières en cas de traitements itératifs (définir la taille des lots en liaison avec les contraintes du dépouillement sur l'ordinateur : taille du fichier spool notamment)

- penser aux documents de suivi du travail (fiches suiveuses), ces documents devant être adaptés à l'organisation retenue pour l'atelier ; définir les cahiers d'enregistrement du travail effectué, les comptes rendus, etc ... ; définir des points de contrôle pour éviter les glissements trop importants par rapport aux prévisions ; etc ...

Chapitre V

La saisie des données

La description technique des moyens de saisie sera faite au fur et à mesure de leur présentation. Dans le corps du chapitre, on indiquera comment la saisie peut interagir avec les problèmes d'organisation de la chaîne de traitement (ateliers manuels ou centre informatique).

Dans cette optique, après avoir défini la saisie des données (étape consistant en la mise sur un support informatique - carte, bandes ou autres supports - de l'information contenue jusque là sur un support plus classique : questionnaires, feuilles de chiffrement, etc ...), on distinguera :

- les moyens classiques de saisie, c'est-à-dire ceux qui isolent la fonction saisie des fonctions amont (collecte, chiffrement) et aval (traitement sur ordinateur : contrôles, redressements),

- les moyens non classiques, c'est-à-dire ceux avec lesquels la saisie des données est un sous-produit d'une autre opération.

A vrai dire, cette distinction, commode d'un point de vue logique, n'est pas toujours réalisée en pratique, certains moyens classiques pouvant être utilisés de façon non classique et vice-versa,

5.1. Les moyens de saisie utilisés de façon classique

a) Le moyen classique le plus connu est la perforation de cartes.
Ses avantages sont :

- une très grande souplesse et sa mise en oeuvre commode (moyen bien connu, utilisable sans investissement important),

- une possibilité de gestion manuelle des fichiers de cartes perforées (mise à jour par substitution d'une carte à une autre, possibilité d'interprétation de l'information perforée, etc ...),

- l'adaptation possible à des problèmes très variés.

./.....

Ses inconvénients sont :

- la limitation à 80 colonnes (nécessité de fractionner l'information contenue sur un même questionnaire) et l'impossibilité de créer des fichiers hiérarchisés : les fichiers de cartes ne peuvent pas refléter la structure du questionnaire, d'où nécessité de procéder ultérieurement à des contrôles de structure lourds et complexes,

- la pauvreté des contrôles possibles (décalage, numériques - non numériques),

- la nécessité d'avoir un matériel de vérification distinct du matériel de saisie,

- les conditions de travail (bruit).

L'utilisation de perforatrices modernes à "buffers" supprime certains de ces inconvénients : suppression du bruit, possibilité de rectification en cours d'enregistrement d'une carte saisie et vérification sur la même machine, extension des contrôles, ...

b) La saisie sur <u>systèmes multiclaviers</u> supprime ces inconvénients. Le système multiclavier est un système composé d'un certain nombre de claviers de saisie (clavier et, en général, écran de visualisation) tous reliés à un mini-ordinateur : les données sont stockées sur un disque d'où elle peuvent être rappelées pour vérification, ou lecture ; une fois vérifiées, les données stockées sur disque sont "vidées" sur bande magnétique. Chaque clavier travaille sous le contrôle d'un programme (généralement appelé "format") qui définit le dessin des enregistrements et les contrôles de validité des données effectués au cours de la saisie.

Les avantages de l'utilisation d'un tel système résident dans :

- la suppression de la contrainte des 80 colonnes qui évite de fractionner arbitrairement le questionnaire entre plusieurs cartes (1 enregistrement = 1 unité logique),

- la possibilité d'enchaîner de façon automatique ou contrôlée des formats divers, ce qui permet d'obtenir un fichier structuré (l'enchaînement logique des différents questionnaires se retrouve au niveau de la saisie),

./.....

- la possibilité d'effectuer certains contrôles de validité des données simultanément à la saisie (attention, cependant, à ne pas trop multiplier ces contrôles : la doctrine à suivre est d'effectuer les contrôles susceptibles de détecter une erreur de saisie, mais pas de chercher à vérifier le chiffrement ou la collecte par des contrôles qui seront effectués ultérieurement, au moment du contrôle informatique).

Les inconvénients du système multiclavier résident, d'une part dans les contraintes de gestion qu'il impose (difficulté d'utilisation en cas de saisie de nombreux fichiers de petit volume) et, d'autre part, dans la nécessité d'une formation supplémentaire donnée aux agents d'encadrement de l'atelier de saisie (manipulation du système, gestion d'une bandothèque, écriture des formats, ...).

Le coût de ces systèmes est fonction du coût du système central (mini-ordinateur, unités de disque, dérouleur de bandes) auquel il faut ajouter le coût marginal des postes de saisie : le seuil de rentabilité (par rapport à la carte perforée) se situe aux environs de 6 postes pour les systèmes peu évolués et d'une dizaine de postes pour les systèmes plus évolués.

c) On n'évoquera que pour mémoire <u>les systèmes mono-claviers</u> qui en raison de leur coût, ne peuvent pratiquement jamais être préférés aux systèmes multiclaviers qui, de plus, ont des possibilités techniques moins étendues. Ces systèmes mono-claviers n'ont, en fait, représentés qu'une étape dans le développement de la technique,

d) En revanche, on assiste au développement de nouveaux supports :

- <u>cassettes</u>,

- <u>disques souples</u> (encore appelés disquettes ou floppy disks).

Les enregistreurs sur ce type de matériel peuvent combiner certains avantages des systèmes multiclaviers - bien qu'en général les logiciels de saisie soient moins évolués - et la souplesse qu'offre la saisie sur cartes, notamment lorsqu'il s'agit de petits fichiers.

Citons pour mémoire un support pratiquement abandonné : le ruban perforé.

./.....

5.2. Les moyens non classiques

a) Dans les procédés dits, de façon impropre, de saisie en ligne (2), la saisie des données est en fait le sous-produit d'une opération de type "atelier manuel" : codification des données par consultation de codes ou de nomenclatures, contrôle de validité ou de cohérence des données. Ces opérations de type "atelier manuel" sont automatisées et sont effectuées sur un terminal de type conversationnel relié à un ordinateur :

- la codification des données est réalisée par consultation de fichiers-informatiques à partir d'informations "entrées" en clair sur le clavier du terminal ; si un article du fichier est identique à l'information entrée, la codification est réalisée ; sinon, on renvoie sur l'écran du terminal la liste des articles du fichier qui ressemblent le plus à l'information entrée,

- les données déjà codifiées du questionnaire (réponses aux questions fermées) sont rentrées par frappe directe du code sur le clavier ; un programme de contrôle des données rentrées déclenche éventuellement l'affichage de messages d'anomalies sur l'écran, messages auxquels l'opérateur du terminal doit réagir.

L'inconvénient d'un tel système est la lourdeur de sa mise en place. surtout si les différents ateliers de saisie en ligne, se trouvent "à distance" de l'ordinateur auquel sont connectés les terminaux. De tels systèmes ne sont donc concevables que pour des applications lourdes ou répétitives, nécessitant la consultation de fichiers volumineux pour la codification des données ou encore des délais de réponse très courts.

b) L'utilisation de mini-ordinateurs peut permettre de résoudre dans certains cas de façon plus souple et moins coûteuse qu'avec des terminaux en ligne des problèmes de nature analogue. Il y a dans ce dernier cas une contrainte apportée par la taille du mini-ordinateur qui limite la souplesse du dialogue opérateur-machine et la taille des fichiers de consultation utilisés pour la codification. Cette solution peut être considérée comme une extension des fonctions d'un système multiclaviers traditionnel.

./.....

---

(2) On préfèrera l'expression de codification en ligne ou de chiffrement en ligne.

La non-universalité des mini-ordinateurs entraînera souvent que ce type de solution n'est envisageable que pour des applications telles qu'il est possible de leur dédier entièrement une machine.

Il est également possible de méler les deux solutions, en utilisant des mini-ordinateurs reliés à un ordinateur plus puissant : si l'essentiel du traitement est effectué sur le mini, le recours au gros ordinateur est toujours possible pour la consultation des fichiers trop volumineux par exemple.

c) Les lecteurs optiques, c'est-à-dire les systèmes qui enregistrent directement sur un support compatible les données issues de la lecture d'informations consistant, soit en marques, soit en caractères mécaniques (machines à écrire, imprimantes) soit en caractères manuscrits (en général, chiffres), sont, en général mal adaptés au traitement d'enquêtes statistiques, dans la mesure où il est souvent impossible d'obtenir des données suffisamment fiables à la source ou des données pouvant être enregistrées telles quelles sans aucune codification.

La nécessité quasi-absolue d'ajouter une étape de codification et de contrôle annihile complètement l'avantage essentiel de cette méthode qui réside en la possibilité de passer directement de la source de l'information au support informatique.

La lecture optique est, en fait, un moyen de saisie bien adapté à certains travaux de gestion (gestion de fichiers administratifs, gestion de commandes, encaissement de factures, ...), ce qui explique son succès, mais son introduction dans des chaînes de traitement statistique s'est souvent soldé par un échec.

On n'évoquera que pour mémoire des procédés dont l'esprit est identique à celui de la lecture optique (mark-sensing, magnéto-lecture) et qui sont presque tombés en désuétude.

5.3. La place de l'atelier de saisie dans la chaîne de traitement

Dans le cas d'une chaîne de traitement linéaire (contrôles/redressements automatiques), la situation géographique de l'atelier de saisie importe peu : si les ateliers manuels et le centre informatique ne sont pas à proximité immédiate, l'atelier de saisie peut être soit centralisé auprès du centre informatique, soit décentralisé auprès des ateliers manuels. Des considérations de rentabilité permettront de retenir une solution.

./.....

Dans le cas de contrôles/redressements itératifs, il importera que les ateliers de saisie se situent à proximité immédiate des ateliers manuels. Dans ce cas, la méthode de transmission des données vers le centre informatique devra être impérativement être étudiée avec soin : il faut éviter, autant que faire se peut, la transmission des données et le retour des messages d'anomalies par la voie postale qui est génératrice de délais en général trop importants ; il convient donc d'adjoindre, si possible, à l'atelier de saisie un terminal de transmission des données par lots (lecteur du support d'enregistrement des données : cartes, bandes, cassettes, et imprimante pour les messages d'anomalies). Selon la disponibilité du matériel existant et le coût des solutions envisageables, cette nécessité de transmission des données peut conduire à la décision de maintenir la saisie sur cartes perforées alors que, pour d'autres raisons, la saisie sur système multiclaviers aurait été jugée préférable.

Il convient enfin, comme pour l'atelier manuel, de ne pas négliger les problèmes de planification et d'ordonnancement des travaux.

Chapitre 5 : Forme et contenu du questionnaire
(J. VAUGELADE Mars 1977)

Le questionnaire est le support matériel de toute étude, il doit d'abord être conçu en fonction de la collecte. Il peut servir en outre au chiffrement et à la saisie (comme le bulletin individuel du recensement français de 1975).

Le questionnaire doit donc répondre à trois objectifs :

- l'enquête
- le chiffrement
- la saisie (on exclut la saisie optique qu'on considère plus appropriée à des opérations administratives qu'aux impératifs d'une enquête).

Il doit donc être discuté et approuvé par tous ceux qui interviennent au long de la chaîne de réalisation (enquêteurs, chiffreurs, saisisseurs), en fonction de deux impératifs :

- faciliter le travail,
- réduire au minimum les risques d'erreur.

Si les discussions font surgir quelques désaccords dans la forme, ils devront être arbitrés en donnant la priorité à l'enquête elle-même.

En effet, pour le chiffrement et la saisie des contrôles et des corrections sont possibles, il en va différemment pour l'enquête, les corrections sont difficiles voire impossibles et toujours coûteuses. On examinera successivement les trois fonctions du questionnaire, l'enquête, le chiffrement, la saisie.

L'enquête

On ne traitera pas du choix des questions en fonction des objectifs de l'enquête, sur ce point le lecteur pourra consulter DESABIE : Théorie et pratique des sondages (DUNOD 1971, pp. 399-468).

Les seuls problèmes à envisager ici sont ceux de l'interaction des autres opérations. Ainsi l'ordre des questions indifférent pour l'informaticien est essentiel pour la collecte. Les questions pouvant indisposer l'enquêté seront plus souvent placées à la fin pour éviter une fin prématurée de l'interrogatoire.

Par contre, pour faciliter le contrôle de la saisie, l'informaticien pourra faire ajouter des questions filtres afin de mieux structurer l'information.

Par exemple, si plusieurs questions ne s'adressent qu'à une partie des enquêtés, il faudra lors de l'exploitation être capable de reconnaître aisément les questionnaires non concernés par ces questions. Ce pourra être le rôle d'une question (ou case de chiffrement) filtre.

Le questionnaire peut être rempli par un enquêteur qui pose des questions à l'enquêté, ou par l'enquêté lui-même.

Dans le premier cas, on peut être tenté de faire réaliser le chiffrement par l'enquêteur. Pour séduisante que soit cette méthode elle est à exclure car trop souvent source d'erreur.

En effet, l'enquêteur doit poser les questions, écouter les réponses et les retranscrire. Si de plus, il doit écrire 1 ou 2 au lieu de M ou F cela fait une opération supplémentaire qui risque de conduire à des erreurs. De plus, il peut être nécessaire de revenir sur une question pour une correction, cela conduirait à des ratures dans les cases de chiffrement et rendrait la saisie plus difficile.

De même cocher la bonne réponse est, pour un enquêteur, une opération répétitive qui peut devenir machinale et donc source d'erreurs. Par contre, quand l'enquêté remplit lui-même le questionnaire, la liste des réponses permet de mieux préciser la question et évite les réponses involontairement fantaisistes. (Aucune technique ne peut éliminer les réponses volontairement fausses). Ceci n'est possible que dans le cas où les réponses possibles sont en nombre limité, ou du moins regroupées en grandes catégories et exclut le cas des questionnaires psychologiques ou sociologiques qui comportent des questions ouvertes d'attitude.

Cocher la bonne réponse est d'ailleurs une instruction insuffisante, il faut être plus précis comme : "Quand des petites cases ont été prévues pour votre réponse, mettez une croix dans celle qui correspond à votre cas", ou "entourez la bonne réponse".

Le chiffrement

Deux cas sont envisageables, le chiffrement est réalisé sur la feuille d'enquête ou bien sur une feuille spéciale.

On verra en annexe un exemple d'imprimé spécial pour la codification.

Le plus souvent, on préfère le chiffrement sur la feuille d'enquête, cela simplifie le travail, il n'y a pas de risques d'omission ou de double chiffrement. Le contrôle en est facilité puisqu'un seul document est manipulé. On trouvera en annexe le questionnaire individuel du recensement français et le questionnaire collectif de l'enquête post-censitaire de Haute-Volta en 1975 qui prévoit une ligne de chiffrement sous chaque ligne d'individu.

Dans tous les cas on préfère que les cases de chiffrement soient distinctes des zones réservées aux réponses. Pour certaines questions quantitatives (comme l'année de naissance) cela entraîne un recopiage qui peut être une source d'erreurs. Mais pour les questions qualitatives et pour les cas spéciaux des questions quantitatives (par exemple les réponses non précisées), les cases de chiffrement laissent au chiffreur la possibilité d'un travail d'interprétation. Cela évite aussi que les surcharges résultant des corrections effectuées à l'enquête ne se trouvent sur les cases de chiffrement, ce qui serait gênant pour la saisie.

Quand le questionnaire présente un aspect collectif comme un questionnaire de ménage, la partie commune de l'identifiant est chiffrée une seule fois, comme dans le questionnaire du recensement voltaïque. Mais s'il y a plusieurs feuilles, cet identifiant devra être répété.

Dans le cas d'une enquête répétitive, comme celle des perspectives de l'industrie, ce qui est déjà connu, adresse des destinataires, produits enquêtés... est imprimé par l'ordinateur d'après les questionnaires précédents (voir questionnaire en annexe).

La saisie

Les contraintes imposées par la saisie sont peu nombreuses mais essentielles. Les cases de chiffrement doivent être groupées, l'ordre doit être évident de haut en bas et de gauche à droite. Les cases doivent être identifiées par un numéro dans le cas de la sai-

sie "classique" sur carte perforée, par des lettres dans le cas de saisie sur multiclavier avec écran de visualisation, de telle façon que la personne qui saisit puisse savoir où elle en est en cas d'interruption ou de doute.

Les constantes pré-imprimées, par exemple le type de carte, seront enregistrées automatiquement. L'identification commune à une série de questionnaires sera saisie une seule fois, elle sera aux mêmes places sur les différents questionnaires.

Conclusions

Le choix de la forme d'un questionnaire est important pour la bonne marche d'une enquête et son dépouillement.

Annexes

- Questionnaire individuel du recensement français de 1975.
- Questionnaire de l'enquête mensuelle sur la situation et la perspective dans l'industrie.
- Grille de chiffrement (enquête migration Mossi).
- Questionnaire collectif de l'enquête post-censitaire voltaïque de 1976.

CHAPITRE 7 - ORGANISATION DES FICHIERS SEQUENTIELS

I - INTRODUCTION

Un fichier est un ensemble d'informations dans lequel on observe des répétitions et des alternatives :

- des répétitions car le fichier est un ensemble de données qui donnent pour les unités statistiques qui y sont décrites les réponses à un ensemble de questions toujours semblables.

- des alternatives dans la mesure, ou d'une unité statistique à l'autre, on peut constater, en dehors des informations communes à toutes, la présence de groupes d'informations facultatives qui n'existent que pour certaines d'entre elles.

Ces deux faits induisent les règles de construction des fichiers que nous nous proposons d'examiner d'un point de vue purement logique d'abord, de celui de leur organisation sur support magnétique ensuite. A ce propos nous nous limiterons aux fichiers séquentiels car la bande magnétique est le support habituel des fichiers d'enquêtes. Dans ce cadre, ainsi limité, les solutions sont encore nombreuses. Nous n'en exposerons qu'une (utilisée par le logiciel de dépouillement d'enquêtes LEDA) qui nous parait la plus cohérente, la plus riche de possibilités, la plus apte à déboucher sur une structure rationnelle des programmes qui auront à traiter les fichiers.

Dans une première partie nous examinerons la répétition qui est le fait le plus important et, nous en déduirons des règles de conduite quant à l'organisation des fichiers sur support séquentiel.

Dans une deuxième nous introduirons la notion d'alternative et nous examinerons ses conséquences sur l'organisation précédemment définie.

Dans une troisième partie enfin nous découvrirons que la structure du fichier induit celle des programmes auxquels il sera appliqué en nous limitant à ceux qui ont un seul fichier en entrée. Le cas des programmes à plusieurs entrées sera examiné dans le chapitre sur la chaîne d'apurement au paragraphe sur la mise à jour.

II - LA REPETITION

II.1 - Exposé du problème

Soit un fichier de "ménages - logements", obtenu par enquête, dans lequel on trouve :

- des données sur chacun des ménages et les logements qu'ils occupent,

- des données sur chacune des pièces de chaque logement,

- des données sur chacun des individus qui constituent chaque ménage,

- des données sur chacun des emplois occupés par chaque individu au cours de l'année précédent l'enquête.

Le fichier contient bien sur plusieurs ménages. Chaque ménage compte un ou plusieurs individus et le logement qu'il occupe compte une ou plusieurs pièces. Enfin chaque individu a, au cours de la période de référence occupé zéro, un ou plusieurs emplois. Nous admettrons toutefois, dans la suite, que le nombre d'occurences de l'individu et de la pièce de logement par rapport au menage peuvent être nulles ce qui sera normal si celui-ci est vide dans le 1er cas, si les données le concernant ne sont pas disponibles dans le 2ème.

On constate plusieurs niveaux de répétition :

- le ménage est répétitif par rapport au fichier,

- l'individu et la pièce de logement sont répétitifs par rapport au ménage,

- L'emploi est répétitif par rapport à l'individu

On peut dire, c'est un problème de terminologie, que le fichier, le ménage, la pièce de logement, l'individu et l'emploi occupé sont des êtres ou unités statistiques de nature et de types différents liés entre eux par des liens hiérarchiques.

.../...

Il importe d'abord de définir la nature des différents êtres. Ainsi :

- Le fichier est une collection de dossiers,

- le ménage est une unité statistique complexe dans sa définition et sa composition puisqu'il réunit les notions de menage proprement dit et de logement et que chacune de ces parties se décompose en unités statistiques plus simples.

- la pièce de logement est une unité statistique simple,

- l'individu est une unité statistique simple en 1ère approche, mais qui devient complexe, en terme de composition, dès l'instant qu'on fait référence à l'emploi qui est lui même une unité statistique simple.

Il importe ensuite de définir les liens hiérarchiques qui les unissent. Ainsi :

- le fichier est l'être majeur, celui qui englobe tous les autres ; nous le dirons de niveau 0.

- le ménage est une unité statistique de niveau 1 ; hiérarchiquement, il est un descendant direct du fichier.

- la pièce de logement et l'individu qui sont des descendants directs du ménage seront de niveau 2 ; Le ménage possède donc deux types de descendants. En outre la pièce de logement, qui n'a pas de descendance, sera dite unité statistique terminale

- l'emploi, descendant direct de l'individu de niveau 2, sera lui même de niveau 3 ; il est également une unité statistique terminale.

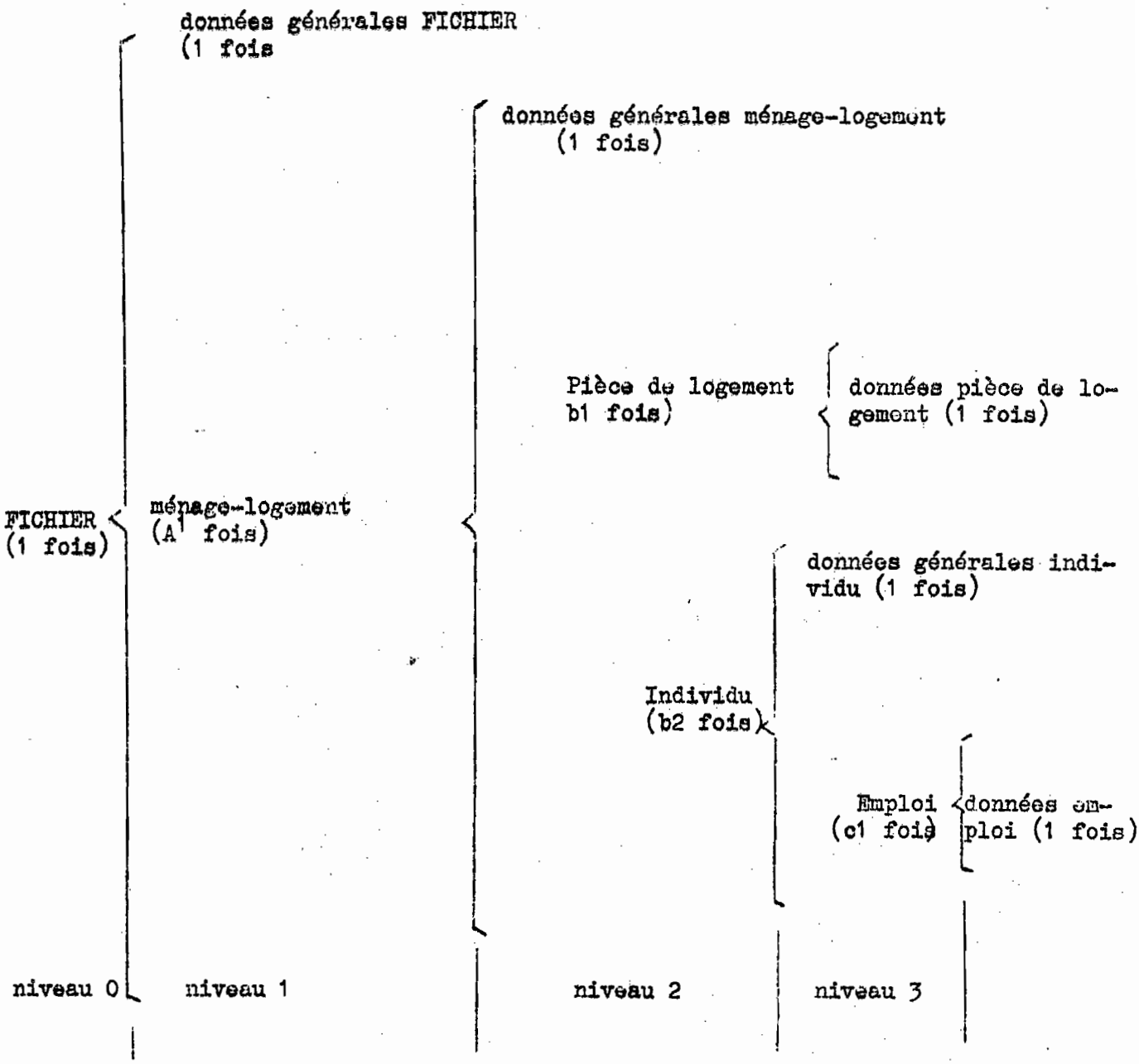
II.2 - Représentation par accolades (LCP-WARNIER)

Schématiquement la structure du fichier que nous venons de décrire pourra se représenter comme suit (1) :

---

(1) schéma mis au point par J. D. WARNIER dans sa logique de construction des programmes (LCP).

.../...



Chaque accolade ou groupe vertical d'accolades définit un niveau :

- le fichier, de niveau 0, existe une fois. On y trouve :

  * une fois les données générales sur le fichier (par exemple son nom, le N° de bande magnétique, le dessin de fichier, etc...).

  * a1 ménages.

.../...

- pour chaque ménage on trouve :

  * une fois les données générales sur le ménage, c'est-à-dire qui ne varient pas pour les unités statistiques de niveau inférieur qui lui sont liées (adresse, type de logement, age du chef, etc...)

  * b1 pièces de logement

  * b2 individus.

- pour chaque pièce de logement on trouve seulement une fois les données "pièce" car il s'agit d'une unité statistique terminal.

- Pour chaque individu on trouve :

  * une fois les données individu

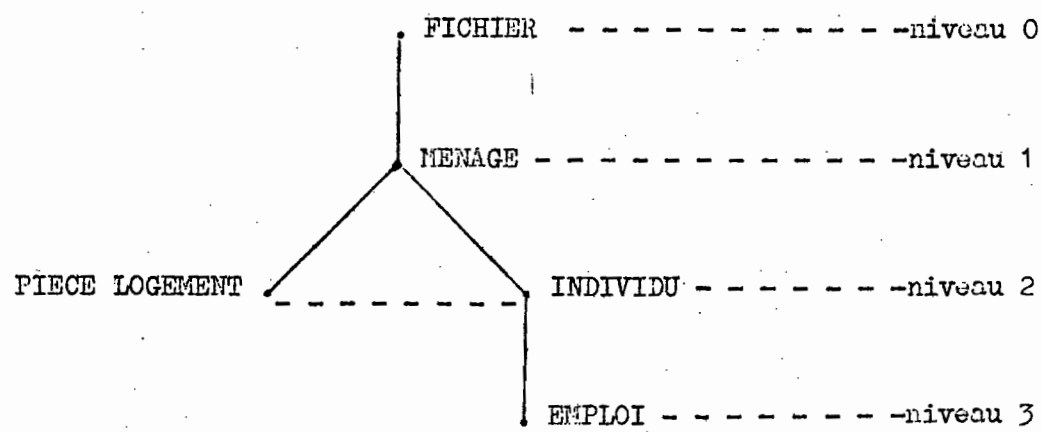
  * c_1 emplois.

- Pour chaque emploi enfin on trouve, comme pour les pièces, les seules données emploi.

a1, b1, b2 et c1 sont des nombres entiers positifs ou nuls. Nous avons en effet admis l'absence de descendance à tous les niveaux. Nous dirons que le ménage, la pièce de logement, l'individu et l'emploi sont des unités statistiques facultatives multiples. Le fichier, quant à lui, est facultatif unique. Il va de soi que l'absence d'une unité statistique, à quelque niveau que ce soit, entraîne ipso facto l'absence de ses descendants des niveaux inférieurs. Si un ménage n'a pas de descendant individu il ne pourra non plus avoir de descendant emplois.

II.3 - Représentation arborescente

La structure du fichier peut aussi être schématisée par un arbre.



.../...

II.4 - Généralisation

La généralisation des notions exposées à partir de l'exemple ci-dessus est assez évidente. Nous énoncerons que :

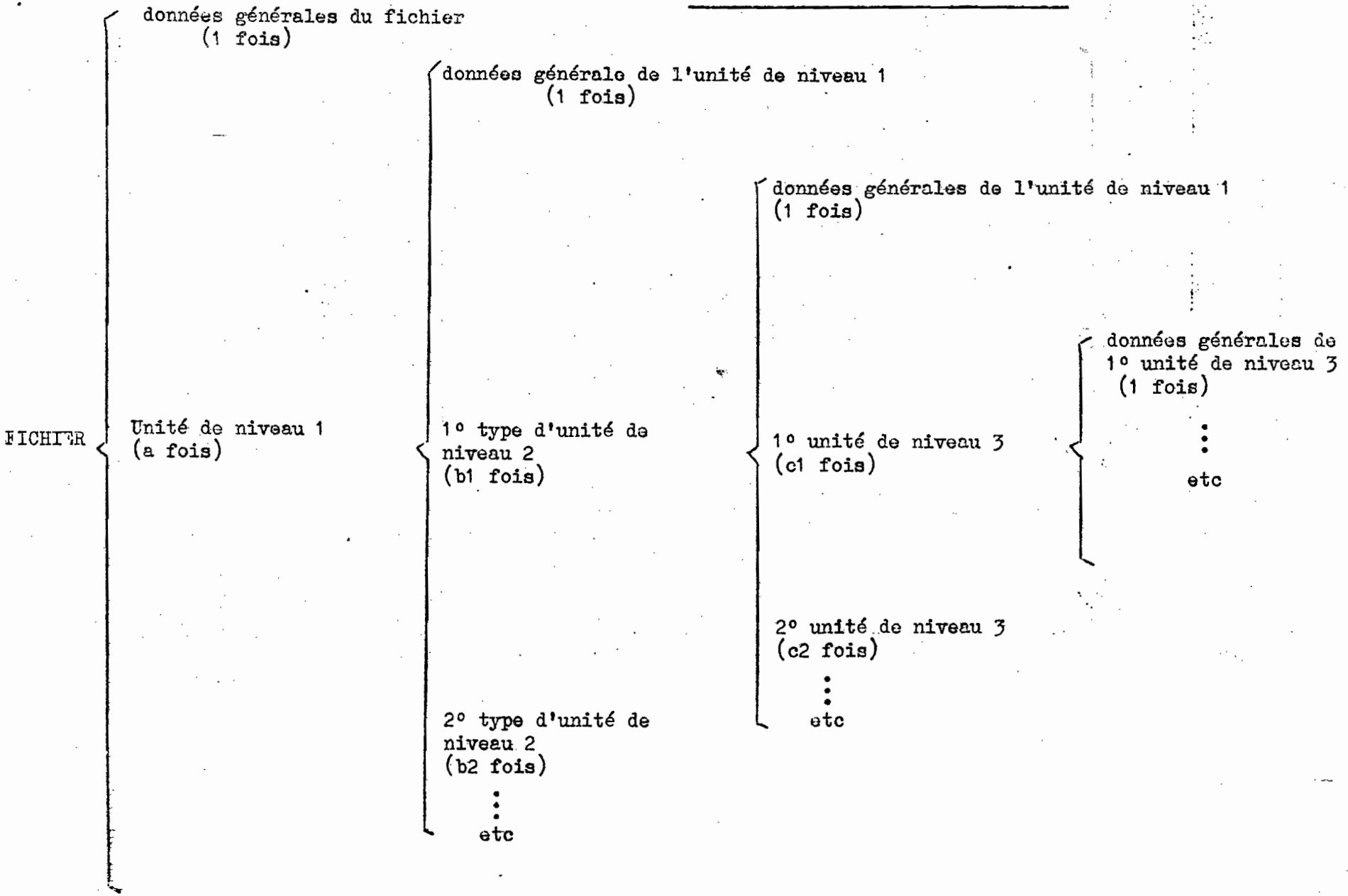
1 - Un fichier peut comporter un nombre quelconque de niveau.

2 - Chaque unité statistique d'un quelconque niveau peut avoir un nombre quelconque de types de descendants. Elle n'en aura éventuellement aucun et sera alors qualifiée d'unité statistique terminale.

3 - On peut faire une exception, non obligatoire, à la règle 2 ci-dessus et dire qu'il ne peut exister qu'un seul type d'unité statistique au niveau 1. On peut en effet considérer que s'il en existe deux (ou davantage) que rien ne relie entre elles, sinon leur présence dans un même fichier on a, de fait, à faire à deux fichiers distincts artificiellement fusionnés. Cette règle parait dictée par le bon sens. Il peut toutefois exister des cas d'espèce qui y font exception. Cependant s'il existe, comme c'est probable un lien, aussi tenu soit-il, entre les différents types d'unités statistiques du niveau 1 on aura peut être intérêt à définir un niveau 1 unique ne contenant que les quelques éléments communs et à rejeter ce qui diffère au niveau inférieur ou dans une structure alternative (voir plus loin).

.../...

GENERALISATION sur SCHEMA VARNIER



II.5 - Organisation des enregistrements - fichiers hiérarchisés

Nous venons de définir des fichiers logiques. Il s'agit maintenant d'envisager comment faire en sorte que la structure logique reste apparente une fois les données transcrites sur un support magnétique à accès séquentiel. Il existe plusieurs façons de structurer un fichier ; nous ne décrirons que celle qui a été retenue pour le logiciel de dépouillement d'enquêtes LEDA.

On définit autant de types d'enregistrements qu'il y a de types d'unités statistique dans le fichier. Chacun ne contient que l'information relative au type d'unité statistique qu'il décrit plus un minimum d'éléments qui décrivent sans ambiguïté les liens hiérarchiques. Le fichier contient toute l'information sans doubles comptes. Les enregistrements sont, en principe, de longueur variable. Dans notre exemple nous auront :

- 1 type d'enregistrement MENAGE,
- 1 type d'enregistrement PIECE DE LOGEMENT,
- 1 type d'enregistrement INDIVIDU,
- 1 type d'enregistrement EMPLOI.

Dès lors il devient nécessaire de disposer d'un code type d'enregistrement qui devra :

a - situer hiérarchiquement chaque type d'enregistrement

b - à un niveau donné différencier les types d'enregistrements.

c - décrire la généalogie d'un type d'enregistrement de niveau inférieur ; il faudra faire apparaître qu'"EMPLOI" est un descendant d'"INDIVIDU" et non de "PIECE".

Il s'agira donc d'un code composé qui dans notre exemple pourra prendre les valeurs suivantes :

- 0 pour le fichier, ce qui signifie qu'on est au niveau 0. Cette indication est suffisante car le fichier n'a pas d'ascendant

- 1 pour le ménage ce qui signifie qu'on est au niveau 1. Là encore cette indication est suffisante car :

.../...

* le ménage étant un descendant direct du fichier, la filiation n'a pas à être indiquée car elle est implicite.

* on a admis qu'il existait un seul type d'unité statistique au niveau 1, il n'est donc pas nécessaire de préciser le type d'enregistrement.

- 21 pour la pièce de logement car :

* l'enregistrement est de niveau 2,

* parmi ceux-ci il est du 1er type,

* il n'y a pas d'indication de la filiation car tous les enregistrements du niveau 2 descendent implicitement du type unique du niveau 1.

- 22 pour l'individu car il s'agit du 2ème type d'enregistrement de niveau 2.

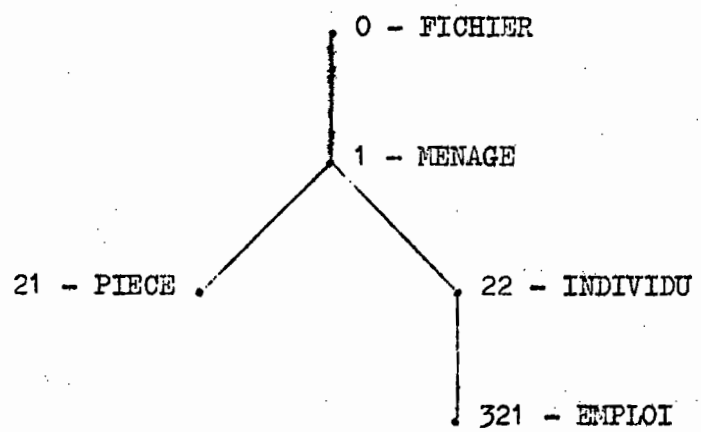
- 321 enfin pour l'emploi :

* 3 car on est au 3ème niveau,

* 2 car l'ascendant de niveau 2 est de type 2 (individu),

* 1 car à l'intérieur du groupe défini ci-dessus par 32, l'enregistrement est de type 1

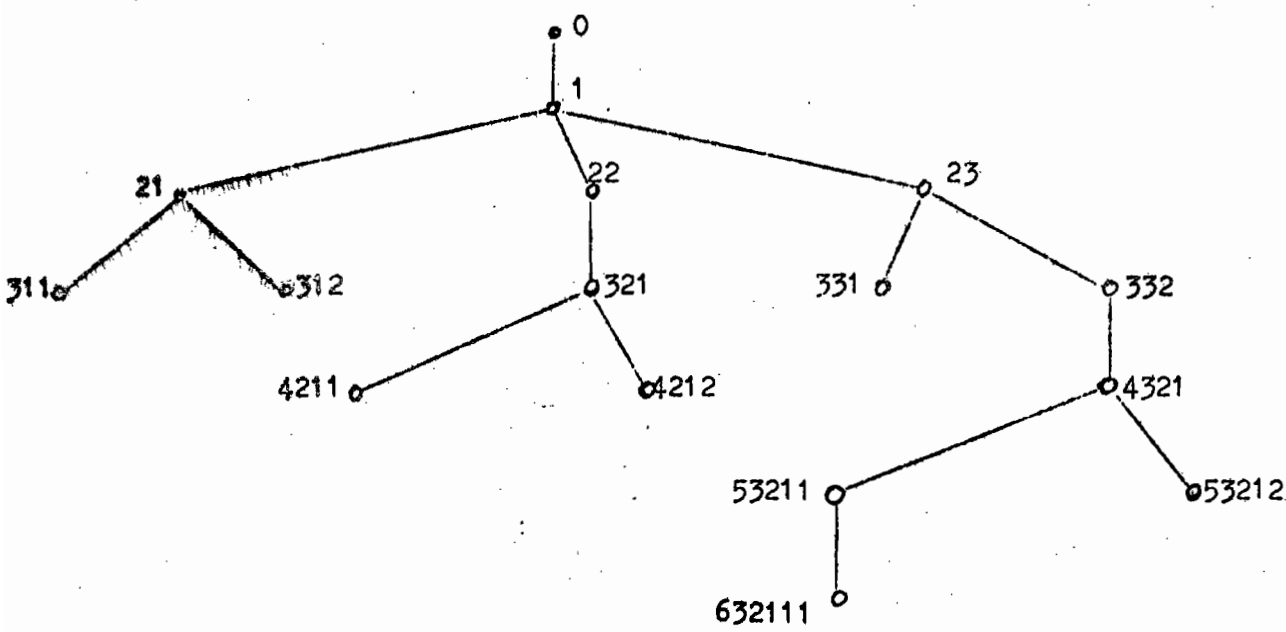
On schématisera comme ceci :



Remarque : L'enregistrement 0 - FICHIER est à priori fictif. Il peut toutefois exister réellement pour des raisons d'harmonie.

.../...

Le procédé étendu à une structure quelconque donnera, par exemple, l'arborescence suivante :



dans laquelle l'enregistrement 632111 est du niveau 6 ; ses ascendants sont :

- l'enregistrement unique de niveau 1 (implicite)
- l'enregistrement niveau 2 de type 3 (63)
- parmi les descendants du précédent l'enregistrement niveau 3 de type 2 (632)
- parmi les descendants du précédent l'enregistrement niveau 4 de type 1 (6321)
- parmi les descendants du précédent l'enregistrement niveau 5 de type 1 (63211),

et dans le groupe ci-dessus définit il est lui-même de type 1(632111)

Au total on pourra énoncer les règles suivantes de construction du type d'enregistrement :

a - un enregistrement de niveau n est identifié par n chiffres.

.../...

b - le 1er chiffre indique le niveau,

c - les n-2 chiffres suivants indiquent la filiation :

- le 2ème rappel l'ascendant de niveau 2
- le 3ème rappel l'ascendant de niveau 3

•
•
•

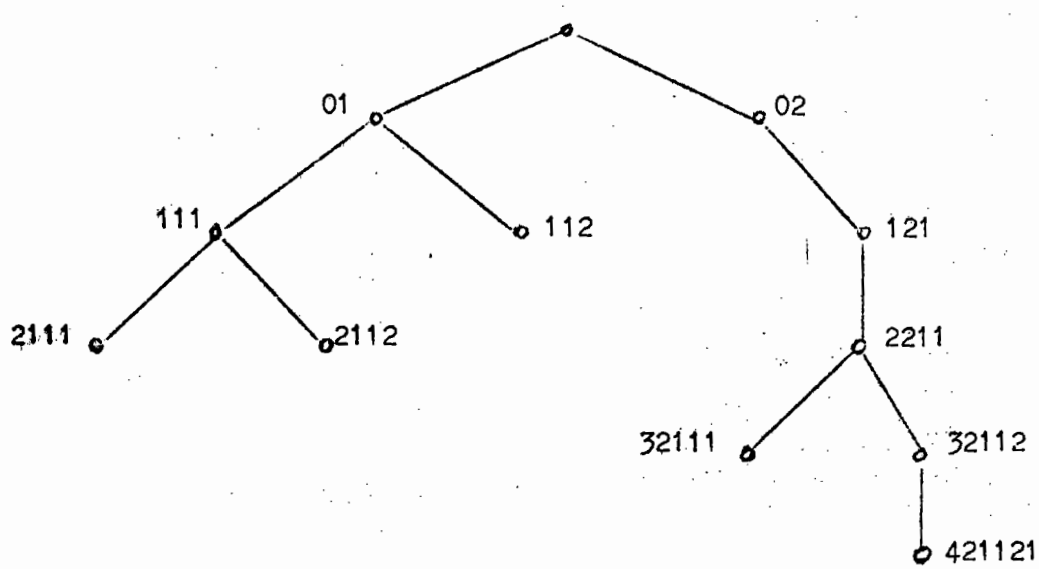
- le n-1ème rappel l'ascendant de niveau n-1

d - le nière chiffre indique le type de l'enregistrement à l'intérieur du groupe défini par les n-1 premiers (niveau et filiation).

e - la règle c ci-dessus n'intervient qu'à partir du 3ème niveau.

f - la règle d ci-dessus n'intervient qu'à partir du 2ème niveau.

Remarque : On peut étendre les règles c et d à tous les niveaux si on admet qu'il peut y avoir plusieurs types d'enregistrements au niveau 1 et (ou) qu'on étend la description à un groupe de fichiers au lieu de la limiter à un seul. On obtiendra alors une description du type suivant :



Les règles devront alors s'énoncer comme suit :

a' - un enregistrement de niveau n est identifié par n+2 chiffres,

b' - le 1er indique le niveau,

c' - les n suivants décrivent sa filiation

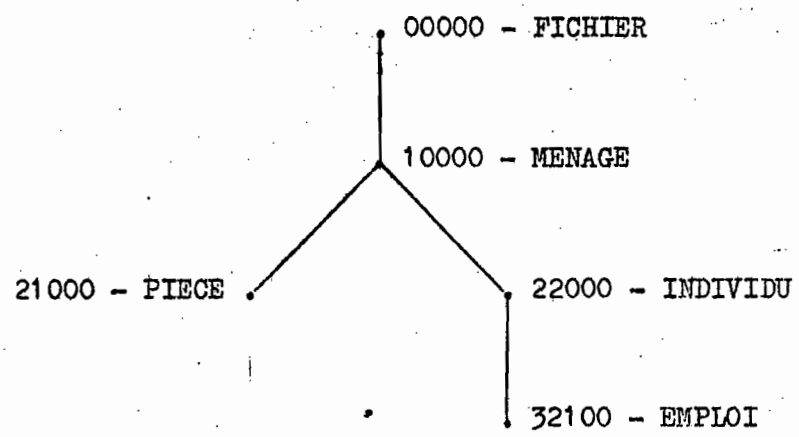
d' - le n+2ème précise son type à l'intérieur du groupe défini par les n+1 premiers.

e' - La règle c' n'intervient qu'à partir du 2ème niveau.

On définira donc le type d'enregistrement sur un nombre de caractères égal au nombre de niveau, non compris le niveau fichier. Pratiquement il sera intéressant, afin d'avoir un code de longueur constante pour tous les fichier d'une application ou d'un groupe d'applications de voir un peu large et de prévoir plus de caractères qu'il ne paraît d'abord nécessaire. Cela permettra en outre d'augmenter le nombre de niveau du fichier sans avoir à en changer le dessin.

Pour les niveaux supérieurs, les positions non significatives du type d'enregistrement seront remplis avec des zéros.

Si on fixe à 5 la longueur du code type d'enregistrement de notre fichier de ménages on obtiendra :



Il est intéressant de déterminer la puissance d'un tel type d'enregistrement ; si sa longueur est n et si on pose qu'il doit être entièrement numérique on aura :

- n niveau d'enregistrement

.../...

- 1 seul type au niveau 1,

- 9 types au niveau 2

- 9^2 types au niveau 3,

.
.
.

- 9^{n-1} types au niveau n

soit :

$$9^0 + 9^1 + \ldots + 9^{n-1} = \frac{9^n - 1}{8} \text{ types possibles}$$

III - L'ALTERNATIVE

Il peut se faire que pour un type d'unité statistique donné des groupes d'informations n'existent que si elle présente certains caractères.

Ainsi dans notre fichier exemple on pourra avoir au niveau du ménage :

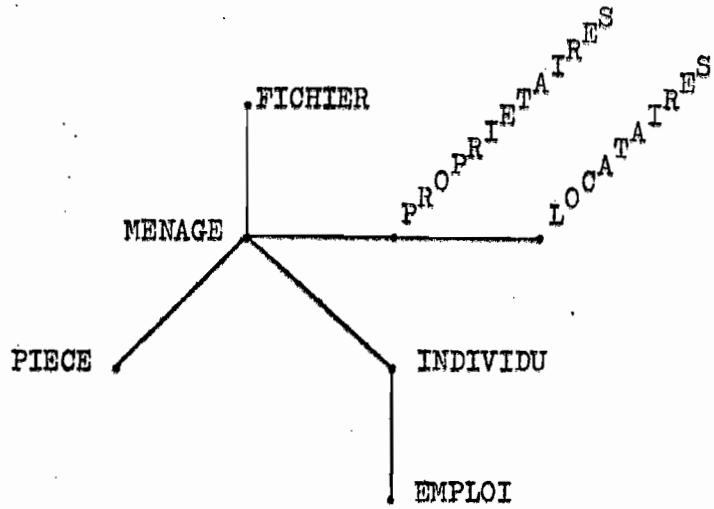
- un groupe d'informations relatives au prix du logement, au mode de financement, à l'endettement, etc... qui ne concernent que ceux qui sont propriétaire de leur logement.

- un groupe d'informations relatives au mode de location, montant du loyer, durée du bail, etc... qui ne concernent que les locataires.

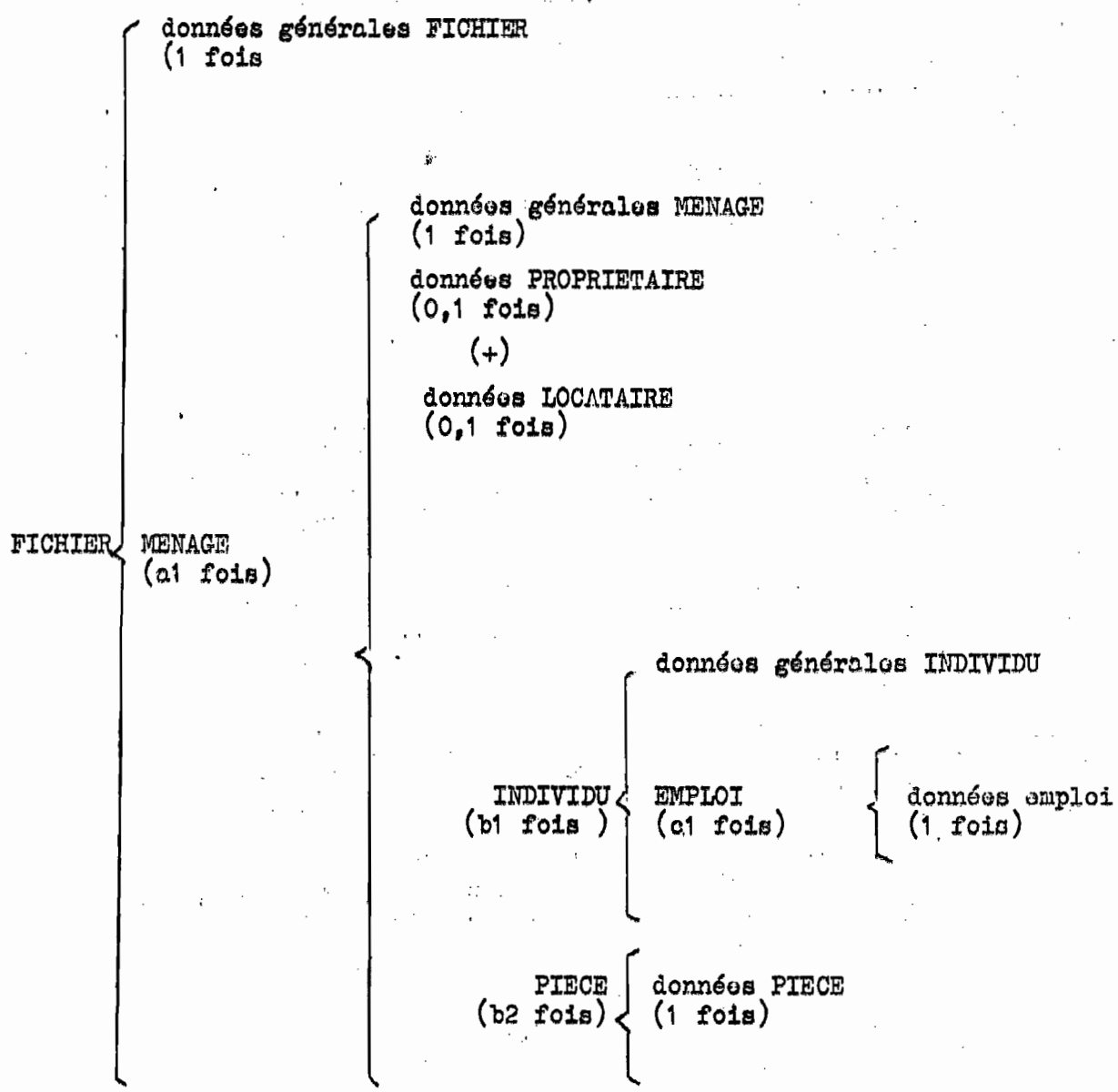
Ces données sont facultatives, puisqu'elles dépendent du statut d'occupation du logement, au même titre que les données individu et pièce de logement. Elles s'en différencient par le fait que, lorsqu'elle existent, elles n'ont qu'une seule occurence possible par ménage. Nous les dirons facultatives unique. Elles ne doivent pas donner lieu à création d'un niveau hiérarchique supplémentaire car elles ne définissent pas un type d'unité statistique distinct du ménage-logement.

Dans la représentation arborescente ce fait se traduira comme ceci :

.../...



Et comme cela dans la représentation par accolades :



La mention "0,1 fois" marque le caractère facultatif unique de ces données.

Le signe (+) (ou exclusif) marque dans ce cas particulier le caractère exclusif des groupes de données qui ne peuvent apparaître simultanément pour un même ménage. Il peut se faire toutefois que des groupes facultatifs uniques ne s'excluent pas. Dans le schéma ci-dessus on les séparera par un signe + (ou inclusif).

Des groupes facultatifs uniques pourront bien sûr être associés à tous les types d'unités statistiques du fichier.

Au niveau de l'organisation physique du fichier, le mieux sera s'ils sont peu nombreux et de faible taille en nombre d'octets, de les intégrer dans les types d'enregistrements auxquels ils se rattachent. La place perdue (pour un enregistrement MENAGE l'une des zones "PROPRIETAIRE" ou "LOCATAIRE" sera toujours vide) sera compensée par un gain de simplicité.

Si par contre ils sont nombreux, de taille importante, et si de plus ils apparaissent assez rarement, il sera souhaitable de les rejeter dans des types d'enregistrements particuliers qu'on baptisera "enregistrements suite" afin de réduire la taille du fichier.

Le type d'enregistrement de ces enregistrements suite sera celui de l'enregistrement de référence auquel on ajoutera un caractère d'identification de l'enregistrement suite. Ainsi :

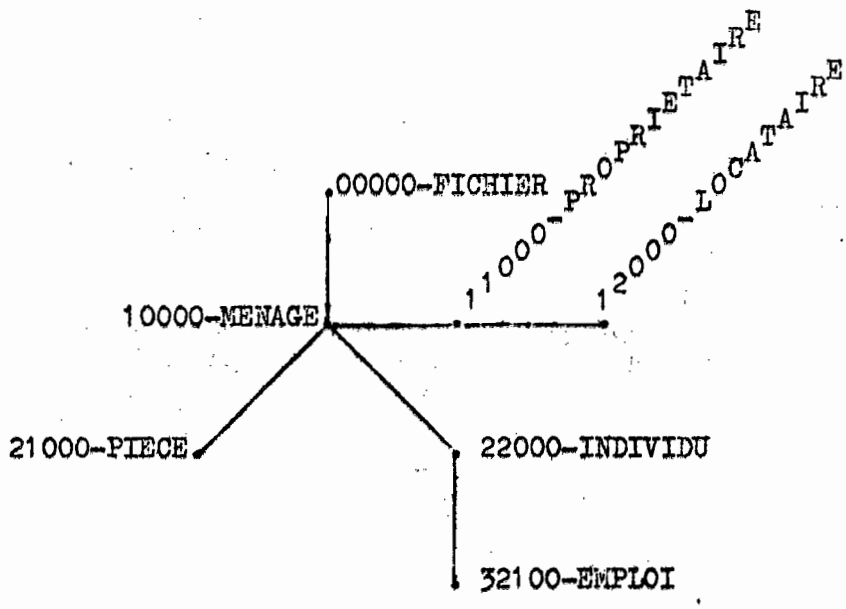
- Les suites d'un enregistrement de type 1 seront identifiées par 11, 12, 13, ..., 19.

- celles d'un enregistrement de type 21 seront identifiées par 211, 212, 213,...., 219.

- celles d'un enregistrement de type 221 seront identifiées par 42211, 42212, 42213, ..., 42219.

Les types d'enregistrements de notre fichier de ménage deviendront :

.../...



Les règles de construction du code type d'enregistrement doivent alors s'énoncer comme ceci :

a - un enregistrement de niveau n est identifié par n chiffres s'il s'agit d'un enregistrement principal et par n+1 chiffres s'il s'agit d'un enregistrement suite.

b - Le 1er chiffre indique le niveau

c - les N-2 suivants décrivent la filiation

d - le nième indique le type d'enregistrement à l'intérieur du groupe défini par les n-1 premiers (niveau et filiation)

e - le n+1ième, quand il existe, indique que l'enregistrement est une suite et précise son type.

f - la règle c ci-dessus n'intervient qu'à partir du 3ème niveau.

g - la règle d ci-dessus n'intervient qu'à partir du 2ème niveau.

.../...

IV - STRUCTURE DES PROGRAMMES TRAITANT UN FICHIER HIERARCHISE

Considérons un programme qui a pour seule entrée notre fichier de ménage hiérarchisé. Nous allons voir que la structure du fichier induit celle du programme et qu'on peut en déduire des règles de construction qui s'appliquent quelque soit la structure du fichier. Nous examinerons successivement le cas ou il n'y a pas d'enregistrements suite et celui ou il y en a.

IV.1 - Fichier sans enregistrement suite

Rappelons que :

1 - Tout type d'enregistrement mineur est facultatif multiple par rapport à son ascendant,

2 - Tout type d'enregistrement mineur est obligatoirement précédé de tous ses ascendants.

Dans le fichier ménage on pourra trouver les références d'enregistrement suivante (chaque enregistrement est figuré par l'indication de son type) :

ménage 1 ! 1 !

ménage 2 ! 1 ! 21 ! 21 !

ménage 3 ! 1 ! 22 ! 321 ! 321 ! 22 ! 22 ! 321 !

ménage 4 ! 1 ! 21 ! 21 ! 22 ! 321 ! 321 ! 22 ! 22 ! 321 !

ménage 5. ! 1 ! 22 ! 321 ! 321 ! 21 ! 22 ! 21 ! 22 ! 321 !

- le ménage 1 est réduit au seul enregistrement ménage

- le ménage 2 comprend en outre deux enregistrements pièce

- le ménage 3 compte 3 individus dont le premier a occupé deux emplois, le second aucun et le 3ème un seul

- le ménage 4 réuni tous les types d'enregistrements

- le ménage 5 aussi mais dans un ordre différent.

.../...

De ce qui précéde on déduit les règles suivantes (à la condition expresse que le fichier soit correctement hiérarchisé) :

a - le 1er enregistrement du fichier est de type 1 (ou 0, si on a défini un enregistrement "fichier" qui sera immédiatement suivi d'un type 1).

b - Un enregistrement de type 1 peut être suivi par :

* un enregistrement de niveau 2,
* un autre enregistrement de type 1,
* rien du tout (fin de fichier)

c - Un enregistrement de type 21 peut être suivi par :

* un autre enregistrement 21,
* un enregistrement 22,
* un enregistrement 1,
* rien du tout (fin de fichier).

d - Un enregistrement de type 22 peut être suivi par :

* un enregistrement 321,
* un enregistrement 21
* un enregistrement 22,
* un enregistrement 1,
* rien du tout (fin de fichier).

e - Les suites possibles d'un enregistrement 321 sont les mêmes que celles d'un enregistrement 22.

Dès lors tout programme ayant ce seul fichier comme entrée pourra avoir une structure générale identique à celle décrite par l'organigramme ci-après dans lequel les tests ont le contenu suivant :

(page 21)

.../...

test 1 : l'enregistrement lu est-il du niveau 2 ?
test 2 : " " " " type 21 ?
test 3 : " " " " type 21 ?
test 4 : " " " " type 321 ?
test 5 : " " " " type 321 ?
test 6 : " " " " type 22 ?
test 7 : " " " " niveau 2 ?
test 8 : " " " " type 1 ?

On remarquera que pour tout type d'enregistrement non terminal le traitement est coupé en deux :

- une partie "début" à la prise en compte de l'enregistrement.

- une partie "fin" après le traitement de tous ses descendants qui ont pu apporter de l'information nécessaire au traitement du père ; cette 2ème partie est éventuellement limitée au seul test sur le type du nouvel enregistrement pris en compte.

Chaque module "traitement début", ou "traitement courant" pour les enregistrements terminaux, se termine par une lecture du fichier d'entrée.

La structure du programme repose sur des tests binaires, ce qui rend nécessaire la présence d'un module complexe (traitement début + traitement fin) "NIVEAU 2". On pourrait supprimer le module simple "début niveau 2" en transformant test 1 en un test ternaire (l'enregistrement est-il du type 21, 22 ou autre ?) mais on ne pourrait éviter la présence du module "fin niveau 2" sans rendre l'algorithme peu comprehensible. Il parait donc préférable de laisser les choses en l'état pour des raisons de symétrie. Si, par contre, il existait 3 types d'enregistrements au niveau 2 (21, 22, 23) il serait tout à fait judicieux que le test 2 soit ternaire (l'enregistrement est-il du type 21, 22 ou 23 ?). De ceci on déduit que lôrsqu'un type d'enregistrement de niveau n a plusieurs types de descendant cela induit la présence dans le programme d'un module complexe "niveau n".

Afin de permettre le bon fonctionnement des tests en fin de fichier il faut associer à la clause fin des ordres de lecture un ordre qui transfert une valeur de padding (999..., HIGH-VALLUE) dans le type d'enregistrement, faute de quoi on "bouclerait" sur le traitement du dernier type d'enregistrement rencontré.

.../...

Ces quelques remarques suffiront à définir, dans le cas général, les régles de construction d'un programme traitant un fichier hiérarchisé quand on aura généralisé comme suit les régles qui définissent l'enchaînement des types d'enregistrements :

a - Le 1er enregistrement du fichier est de type 1 (ou 0, si...):

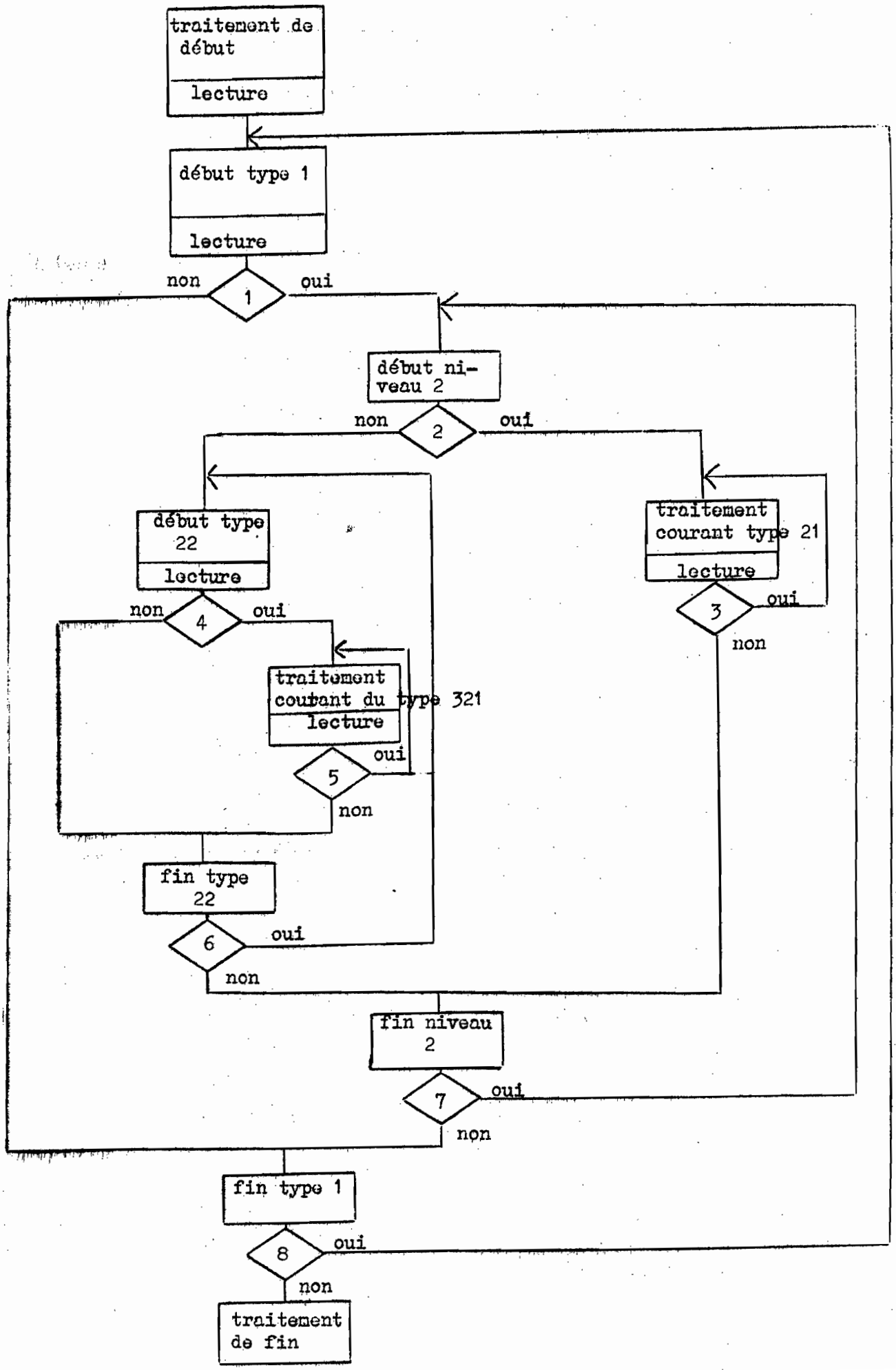
b - Un enregistrement de type 1 peut être suivi par :

* un autre enregistrement de type 1,
* un enregistrement de niveau 2,
* rien du tout (fin de fichier)

c - Un enregistrement de niveau n est suivi par :

* un autre enregistrement de niveau n descendant du même enregistrement de niveau n-1 que lui,
* s'il est non terminal, un enregistrement de niveau n+1 qui appartient à sa descendance,
* un enregistrement placé plus haut que lui dans la hiérarchie,
* rien du tout (fin de fichier

.../...


traitement de début
lecture
début type 1
lecture
non
1
oui
début niveau 2
non
2
oui
début type 22
lecture
traitement courant type 21
lecture
non
4
oui
3
oui
non
traitement courant du type 321
lecture
oui
5
non
fin type 22
6
oui
non
fin niveau 2
7
oui
non
fin type 1
8
oui
non
traitement de fin

IV.2 - Fichiers munis d'enregistrement suite

Dans notre fichier ménage les séquences d'enregistrements possibles sont alors les suivants :

! 1 !

! 1 ! 11 !

! 1 ! 12 !

! 1 ! 11 ! 21 ! 21 !

! 1 ! 12 ! 22 ! 321 ! 321 ! 22 ! 321 !

! 1 ! 11 ! 21 ! 22 ! 22 ! 321 ! 22 ! 22 ! 321 !

etc...

Les règles d'enchaînement des types d'enregistrements deviennent :

a - idem

b - un enregistrement de type 1 peut être suivi par :

* un seul enregistrement de sa suite si ses suites s'excluent, plusieurs si elles ne s'excluent pas ,

* les autres suites possibles restent les mêmes

c - un enregistrement de type n peut être suivi par :

* un seul enregistrement de sa suite si ses suites s'excluent, plusieurs si elles ne s'excluent pas,

* les autres suites possibles restent les mêmes.

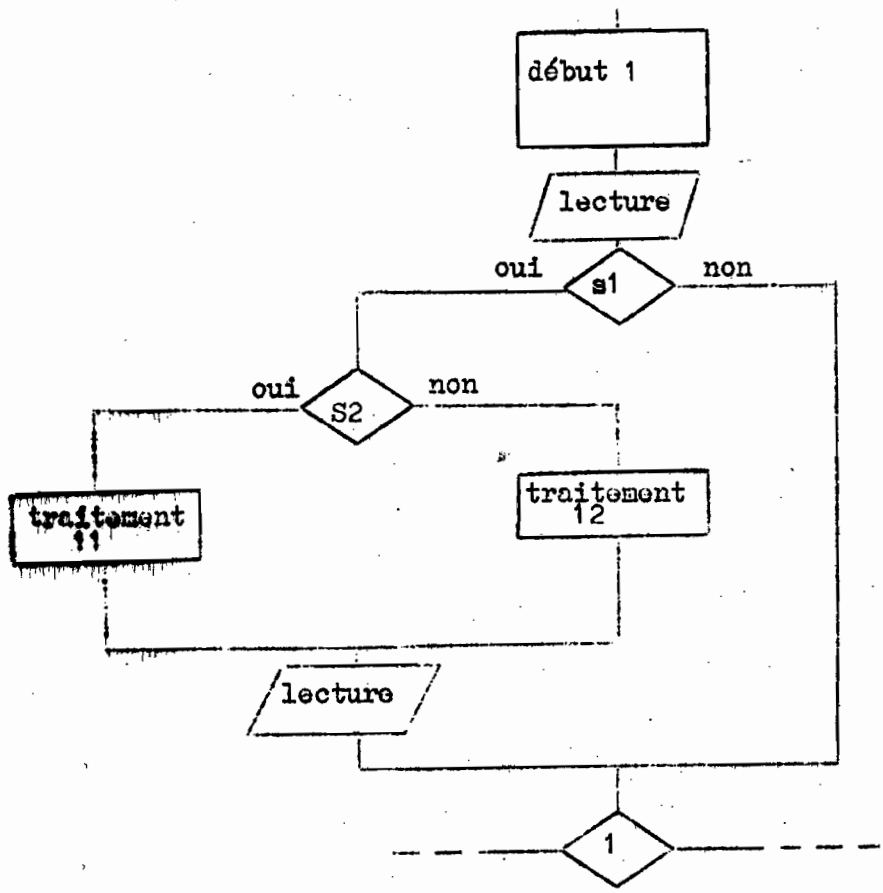
Quant à la structure des programmes elle restera pour l'essentiel ce qu'elle était lorsqu'on traitait des fichiers sans enregistrements suite. Seuls seront affecté les modules "début" des types d'enregistrements munis de suites.

En effet celles-ci font partie de l'unité statistique décrite pour l'essentiel dans l'enregistrement principal, mais l'information est ici répartie sur plusieurs enregistrements au lieu d'être concentrée dans un seul. Une fois qu'elle aura été toute entière prise en compte on se retrouvera dans une situation identique à celle qu'on avait après le module début dans le cas précédent. Nous distinguerons le cas ou les suites s'excluent de celui ou elles peuvent cohabiter en l'illustrant par l'exemple du fichier ménage.

.../...

IV.2.1 - Suites exclusives

Les suites sont associées à l'enregistrement de 1 ; c'est donc le module "début 1" qui va se trouver modifié



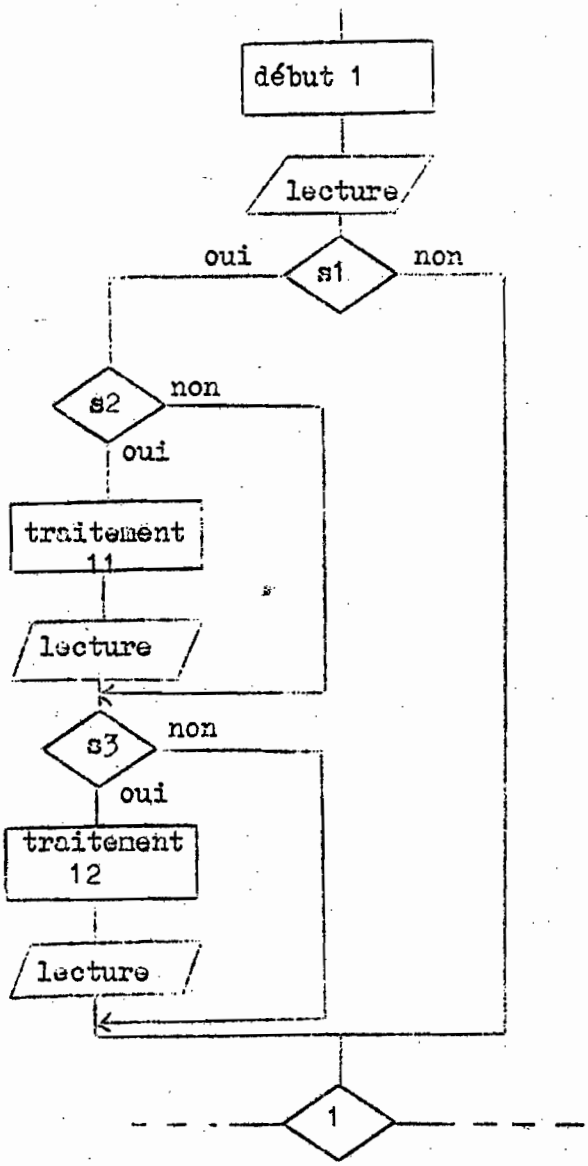
test 11 : L'enregistrement lu est-il un enregistrement sui (autrement dit est-il de niveau 1 et le 2ème chiffr type d'enregistrement est-il ≠ 0 ?)

test 12 : L'enregistrement est-il de type 11 ou 12 ?

Après le traitement 11 ou le traitement 12, on l n nouvel enregistrement qui ne peut pas être une suit on retrouve donc le test 1 déjà connu et le programme tinu de dérouler comme précédemment

/...

IV.2.2 Suite inclusives



test 11 : s'agit-il d'un enregistrement suite

test 12 : L'enregistrement est-il de type 11 ?

test 13 : L'enregistrement est-il de type 12 ?

Chapitre 8

La chaine d'apurement

Rédacteur : L. BREAS

Etat de la rédaction : définitive

Plan du Chapitre :

- Introduction,
- Les fonctions d'entrée - sortie,
- Edition des messages d'erreurs et d'anomalies,
- Tri et fusion. Le tri par groupes,
- Production d'indicateurs sur l'état du fichier,
- Contrôles statistiques - Histogrammes,
- Contrôles et redressements individuels,
- La mise à jour,
- Le pòlissage,
- Synthèse,
- Cönclusion.

LA CHAINE D'APUREMENT

I - INTRODUCTION

II - LES FONCTIONS D'ENTREES - SORTIES

II-1- Un exemple de programme de carte-à-bande : INTRO

II-1-1- Les cartes REPERE
II-1-2- La carte COMPTAGE

II-2- Un exemple de programme d'impression : GUTENBERG

II-2-1- Principe de la mise en page
II-2-2- Gestion des interruptions et des reprises
II-2-3- Possibilités complémentaires

III - EDITION DES MESSAGES D'ERREURS ET D'ANOMALIES

III-1- Caractéristiques des messages d'anomalies

III-2- Justification du programme d'édition des messages

III-3- Généralisation du programme d'édition des messages

III-3-1- Edition fruste
III-3-2- Edition élaborée

IV - TRI ET FUSION - LE TRI PAR GROUPES

IV-1- Principe du tri par groupes

IV-2- Solutions

V - PRODUCTION D'INDICATEURS SUR L'ETAT DU FICHIER

V-1- Les informations utiles

V-2- Utilisation des informations

V-2-1- Contrôle de qualité
V-2-2- Contrôle d'exhaustivité
V-2-3- Correction des contrôles et redressements

V-3- Solutions techniques

VI - CONTROLES STATISTIQUES - HISTOGRAMMES

VII - CONTROLES ET REDRESSEMENTS INDIVIDUELS

VII-1- Contrôle à redressement manuel et contrôles à redressement automatiques

VII-2- Les principaux types de contrôles individuels

VII-2-1- Les contrôles de la mise à jour
VII-2-2- Les contrôles de structure

VII-2-2-1- Présentation et délimitation du champ
VII-2-2-2- Méthodologie du contrôle de structure

VII-2-2-2-1- Structure du fichier
VII-2-2-2-2- Tri du fichier
VII-2-2-2-3- Le contrôle

VII-2-3- Les contrôles de validité et cohérence des données
VII-2-4- Analyse et amélioration des contrôles et redressements

VII-2-4-1- Analyse des évènements
VII-2-4-2- Contrôle des programmes

.../...

.../...

I - INTRODUCTION

La chaîne d'apurement a pour but d'assurer la prise en compte des données et les contrôles et redressements manuels et automatiques qui, partant des données brutes saisies, conduisent à un fichier jugé propre et exhaustif, donc susceptible d'être soumis à une exploitation statistique.

Elle devra par ailleurs produire des indicateurs qui permettent de :

- contrôler à tout moment l'avancement du travail,

- déterminer la qualité finale du fichier,

- déterminer les modifications qu'il faudrait apporter aux contrôles et redressements pour améliorer la qualité à l'occasion d'une enquête ultérieure.

La chaîne d'apurement est la partie la plus délicate du système de dépouillement informatique d'une enquête. En effet c'est là que les risques d'erreurs sont les plus grands :

- erreurs d'analyse dans la définition des contrôles et redressements à opérer dus à une connaissance insuffisante du milieu enquêté qui conduit à mal préjuger des réponses qui seront apportées aux questions posées et à prévoir des contrôles inutiles d'une part, à en négliger d'autres qui seraient utiles d'autre part,

- erreurs dans l'enchaînement des contrôles et redressements qui peut conduire à modifier une variable originellement correcte après un contrôle croisé avec une variable erronée,

- risque d'erreurs dans l'analyse organique et la programmation plus grand qu'au stade de la codification ou de la tabulation du fichier dans la mesure où les traitements sont souvent plus complexes et la détection des erreurs plus difficile du fait de la mauvaise connaissance des résultats qu'on va obtenir.

.../...

De plus la chaîne d'apurement reçcit une information d'origine humaine (en provenance des ateliers de gestion et de saisie) et, lorsqu'on utilise les redressements manuels retourne au premier, sous forme de messages d'erreurs et anomalies, une information qui a du être préalablement "humanisée". Il est dès lors nécessaire de coordonner les actions des équipes informatique et de gestion manuelle, de planifier leurs tâches, d'organiser le dialogue en levant en particulier les barrières que dressent les jargons techniques des uns et des autres afin d'éviter les incompréhensions (des messages de la part des gestionnaires, des corrections de la part des informaticiens et de leurs programmes), d'apprécier au plus juste la capacité de traitement des uns et des autres afin d'éviter les pointes de charge et les périodes creuses.

La réalisation des objectifs impartis à la chaîne d'apurement suppose qu'on fasse intervenir des fonctions de :

1 - prise en compte des données (carte-à-bande,...),

2 - contrôles,

3 - production des messages d'erreurs ou d'anomalies constatés par les contrôles,

4 - correction ou redressement,

5 - mise à jour pour la correction ou la suppression des données erronées et l'introduction de données nouvelles,

6 - production d'indicateurs sur l'état du fichier : comptages d'erreurs, comptage des effectifs du fichier, tableaux de contrôle de la structure statistique du fichier, etc...,

7 - édition des messages et indicateurs,

8 - classement car l'ordre dans lequel les US sont soumises à contrôle et redressement, de même que l'ordre dans lequel doivent être édités les messages et indicateurs, n'est généralement pas indifférent,

9 - fusion pour regrouper des sous-ensembles de données traitées séparément,

10 - polissage ou effacement des incohérences résiduelles après l'apurement.

.../...

Ces fonctions peuvent être analysées de 3 points de vue :

1 - du point de vue de leur généralité : certaines d'entre elles sont entièrement spécifique de chaque enquête (contrôles, redressements), d'autres s'appliquent à toutes au prix d'un paramètrage peu important (tri).

2 - du point de vue des ressources machines qu'elles requièrent l'objectif est là d'utiliser au mieux les ressources de l'ordinateur,

3 - de celui enfin de leur enchaînement logique dans la chaîne d'apurement.

Cette triple analyse appliquée à chacune des fonctions permettent de déterminer son degré de généralité et son degré d'autonomie par rapport aux autres. Partant de l'idée qu'on doit trouver dans une chaîne d'apurement d'enquête un certain nombre de problèmes types auxquels on peut apporter des solutions types on pourra aboutir pour certaines fonctions soit à la définition de programmes standards utilisables dans toutes les enquêtes, soit à des algorithmes transposables d'une enquête à l'autre au prix de modifications assez minimes. Une telle démarche présente le triple intérêt de réduire les coûts d'analyse et programmation une fois réalisé l'investissement de départ, de réduire voire d'annuler les risques d'erreurs dans la solution des problèmes traités par ces programmes et algorithmes standardisés, enfin d'aboutir à une certaine standardisation des supports d'informations échangés avec le statisticien (messages d'anomalies, listings de contrôles, etc...) qui à la longue devrait grandement faciliter le dialogue.

* *

*

Nous allons dans la suite de ce chapitre faire un exposé aussi exhaustif que possible de toutes les fonctions qu'on peut prendre en compte dans l'apurement d'un fichier d'enquête. Les réaliser toutes pour une enquête donnée représente une somme de travail très importante. Il faut bien voir que toutes les enquêtes ne justifient pas un tel luxe de précautions. Les enquêtes simples, portant sur des effectifs enquêtés peu nombreux, non répétives peuvent se contenter d'un apurement plus fruste. Il appartiendra au responsable après avoir apprécié le coût de chacune et le gain de qualité ou de sécurité attendu de choisir celles qui lui sont nécessaires.

La chaîne d'apurement telle qu'elle est décrite ici traite des fichiers séquentiels en traitement par lots sur gros ou moyen ordinateur (au moins 120 K-octets de mémoire). Nous avons retenu ce cas de figure parce qu'il nous semble que c'est celui qui doit se présenter le plus communément. Il serait toutefois judicieux de se demander en quoi son organisation se trouverait modifier par l'utilisation d'un petit ordinateur, par des contrôles intéractifs, par l'utilisation de fichiers en accès direct. Il serait prématuré de tenter de répondre à ces questions à ce stade du travail. Nous y reviendrons à la fin du chapitre.

.../...

II - LES FONCTIONS D'ENTREES - SORTIES

Ce sont les fonctions carte à support magnétique et support magnétique à imprimante. La première a pour but de transcrire une information d'origine humaine, stockée sur un support directement lisible par l'homme, la carte perforée, sur un support magnétique bande ou disque, que seul l'ordinateur saura interpréter. La seconde devra à l'inverse transcrire une information élaborée par l'ordinateur d'un support magnétique sur un support lisible par l'homme, le papier.

A leur propos on peut faire les remarques suivantes :

1 - Les organes d'entrées - sorties sur des supports papier (cartes perforées ou listings) sont très lents comparées aux organes de calcul et d'entrées - sorties sur supports magnétiques. Les performances des matériels d'entrées - sorties les plus fréquemment utilisés sont grosso modo les suivantes :

- 800 caractères/seconde pour un lecteur de carte à 600 cartes/minute,
- 4500 caractères/seconde pour une imprimante à 2000 lignes minute,
- 320000 caractères/seconde pour un lecteur de bande à 320 K octets/seconde.

La différence entre les performances des périphériques lents et celles des périphériques rapides est donc considérable.

Dans un environnement de multiprogrammation la taille mémoire est fixée en fonction de la taille moyenne des programmes, de la vitesse des organes d'E/S sur support magnétique et de celle de l'UC de telle manière que la mémoire puisse accueillir un nombre de programmes tel que la somme de leurs utilisations de l'UC sur un intervalle de temps donnée soit aussi élevé que possible. Si un ou plusieurs programmes utilisent de manière massive une entrée sur carte ou une sortie à l'imprimante le rapport temps d'entrée - sortie sur temps UC va devenir si important que l'objectif ci-dessus ne pourra être atteint.

Pour maintenir un taux d'utilisation de l'UC satisfaisant, la seule solution consiste à affectar à un tel programme des ressources en taille-mémoire et périphériques rapides si réduites que le nombre de programmes qui pourront être activés en même temps que lui n'en soit pratiquement pas changé.

.../...

2 - La lecture d'un fichier sur carte ou une impression peuvent durer très longtemps dès l'instant que les volumes sont importants. Il faudra par exemple 15 heures pour lire 500.000 carte à la vitesse théorique de 600 cartes/minute et plus probablement 20 heures à une vitesse pratique de 400 cartes/minute. Une telle saisie devra généralement être fractionnée en plusieurs lots.

Par ailleurs les risques d'incidents sont importants (bourrages de cartes ou de listings, bacs de cartes oubliés ou saisis en double, papier mal réglé, etc...) qui obligent à stoper puis à reprendre le traitement. Il faut donc, si on veut éviter à chaque fois de reprendre les opérations au début, associer à ces fonctions des procédures d'interruption et de reprise très efficaces.

3 - D'un point de vue fonctionnel il est souvent nécessaire de disjoindre ces fonctions des traitements qui les suivent ou les précèdent. Ce sera quasi systématiquement le cas pour la saisie de cartes car ces dernières devront être triées avant d'être soumises à tout autre traitement.

Ces considérations conduisent à isoler le carte-à-bande et l'impression dans des programmes de service qui :

- mobilisent un minimum de ressources des unités rapides (20 K octets de mémoire et un périphérique rapide). Ils devront pour cela être écrits en langage assembleur,

- gèrent les interruptions et les reprises,

- soient généralisés à toutes les applications, ce qui est aisément réalisables car les problèmes restent très similaires d'une application à l'autre.

A titre d'exemple on peut présenter les caractéristiques essentiels des programmes de carte-à-bande et d'impression différée utilisés à l'INSEE (INTRO et GUTENBERG).

.../...

II-1- Un exemple de programme de carte à bande : INTRO

Le programme est écrit en assembleur. Il utilise de 15 à 20 K octets de mémoire.

Il permet la reprise du carte-à-bande après interruption et le comptage des cartes selon des critères de ventilation simples au moyens de cartes spéciales ajoutées aux cartes de données qu'on appelle cartes REPERE et carte COMPTAGE.

II-1-1- Les cartes REPERE

Ce sont des cartes qui sont insérés dans le flot de cartes données à intervalles réguliers (une en tête de chaque bac par exemple soit approximativement toutes les 2000 cartes). Elles ont le dessin suivant :

Col 1 à 20 : non utilisées,

Col 21 à 60 : intitulé libre. Titre de l'enquête par exemple,

Col 61 à 76 : la mention 'CARTE.REPERE.NØ',

Col 77 à 79 : le n° de la carte repère.

Les cartes repères seront numérotées séquentiellement de 1 à n.

La structure de la carte repère, avec les colonnes 1 à 20 vides en particulier, permet d'éviter qu'elle ne soit confondue avec les cartes données.

Le programme vérifiera que les nos de cartes reprises s'enchaînent correctement (sans omission ni double) et signalera toute anomalie. On réduira ainsi les risques d'oubli et de double saisie.

Lorsqu'on voudra reprendre le carte-à-bande après interruption on fournira au programme, par paramètre, le n° de la carte repère à partir de laquelle on veut reprendre. Le programme viendra alors ajouter sur la bande de saisie les nouvelles cartes à saisir à la suite de celles qui s'y trouvent déjà.

.../...

II-1-2- La carte COMPTAGE

Elle est placée en tête du flot de carte à saisir. On y décrit sous la forme position-longueur les codes selon lesquels on veut ventiler les comptages.

Une carte comptage ainsi libellée

1	21	71	80
	* DEPARTEMENT-COMMUNE * (5, 2, 8, 3)	CARTE COMPTAGE	

permettra d'avoir des comptages par départements et communes.

L'emploi de la carte COMPTAGE qui est facultatif suppose que toutes les cartes du flot de saisie soient de même type.

Il y a par ailleurs comptage automatique du total des cartes saisies et des cartes saisies entre deux cartes REPERE.

Remarque :

Les cartes REPERE sont incluses au fichier des données qu'elles parasitent pour la suite des traitements. Il faudra donc les éliminer dès l'étape suivante dans le tri qui suit normalement le carte-à-bande (ce pourra être une fonction annexe du tri par groupe - voir paragraphe III) ou le 1er programme de contrôle. Une autre solution consisterait à gérer les cartes repère sur un petit fichier disque distinct du fichier des données.

II-2- Un exemple de programme d'impression : GUTENBERG

Le programme est écrit en assembleur. Il utilise 15 à 20 KK octets de mémoire.

.../...

II-2-1- Principe de la mise en page

Le programme GUTENBERG est conçu pour des imprimantes à 132 caractères par ligne. Il reçoit des enregistrements de 135 caractères composés par le programme chargé de l'édition :

- les 132 premiers sont une image de la ligne à imprimer.

- les 3 derniers indiquent les sauts de pages et de lignes qui doivent précéder ou suivre la ligne imprimée :

* le 133ème indique le saut de page avant impression ; il prend la valeur 0 lorsque la ligne doit être imprimée sur la même page que la précédente et la valeur 1 lorsqu'elle doit être imprimée en tête de la page suivante.

* le 134ème indique les sauts de lignes avant impression il prend les valeurs 0 à 9 selon qu'il faut laisser 0 à 9 interlignes entre la ligne précédente et la ligne imprimée.

* le 135ème indique les sauts de lignes après impression il prend les valeurs 0 à 9 selon qu'il faut laisser 0 à 9 interlignes entre la ligne imprimée et la ligne suivante.

Les 3 sont évidemment combinables. Ainsi la valeur 199 en position 133 à 135 :

- provoquera un saut de page,
- ménagera 9 lignes vides en têtes de la nouvelle page,
- imprimera la ligne,
- ménagera 9 lignes vides après.

II-2-2- Gestion des interruptions et des reprises

GUTENBERG se réserve en tête de chaque page une ligne sur laquelle il imprime le n° de page et le n° de la première ligne de la page contenant de l'information qui est également le n° d'ordre de l'enregistrement dans le fichier d'édition. Lorsqu'une reprise d'impression est nécessaire on indique au programme, par paramètre, le n° de la ligne à partir de laquelle on veut reprendre après l'avoir relevé sur le listing déjà imprimé. Il reprendra alors la lecture du fichier d'impression à son début mais ne commencera l'impression qu'à partir de la ligne indiquée.

.../...

II-2-3- Possibilités complémentaires

GUTENBERG imprime au départ un cadre de réglage qui permet de régler l'imprimante. Comme il est activé au pupitre on pourra en demander l'impression plusieurs fois de suite jusqu'à ce que le réglage soit jugé parfaitement satisfaisant et qu'on puisse lancer l'impression du fichier proprement dit.

Il imprime par défaut 50 lignes par page avec des sauts de pages automatiques indépendants de ceux qui sont commandés par le 133ème caractère. Ce nombre de lignes peut être modifié par paramètre d'une part ; la fonction peut être annulée de telle manière que tous les sauts de lignes soient commandés par le 133ème caractère seulement d'autre part.

../...

III - EDITION DES MESSAGES D'ERREURS ET D'ANOMALIES

III-1- Caractéristiques des messages d'anomalies

Un message d'anomalie bien conçu devra présenter les caractéristiques suivantes :

1 - Identifier sans ambiguité l'US en anomalie, c'est à dire contenir son identifiant bien mis en valeur.

2 - Situer le message dans l'espace et dans le temps, c'est à dire préciser la date d'édition, le n° d'ordre du tour de contrôle, les caractéristiques de la strate à laquelle appartient l'US qui sont aussi les critères de tri des messages d'anomalies (cf point 3 ci-après) qui permettront de réaliser la ventilation entre et au sein des services correcteurs (exemple : s'il existe un service correcteur par région et à l'intérieur des services, une répartition du travail par département et commune entre les équipes on retiendra la région, le département et la commune comme critères de situation dans l'espace.

3 - Etre triable avant impression afin de faciliter la répartition entre services (cf point 2 ci-dessus).

4 - Les textes doivent être clairs et modifiables, il importe qu'ils soient bien compris par les gestionnaires qui auront à examiner les erreurs et anomalies. Ce n'est que par tatonnements qu'on aboutira à des textes satisfaisants pour tous. Il faut donc se réserver la possibilité de les modifier aisément.

5 - Contenir l'information telle qu'elle a été saisie, c'est à dire telle qu'elle figure sur le questionnaire (sauf bien sur en cas d'erreur de saisie) auquel le gestionnaire aura constamment recours pour corriger l'erreur.

6 - Facile à manier pour le gestionnaire, celui-ci a pour habitude d'ouvrir un dossier par enquête dans lequel il versera d'abord le questionnaire et, s'il y a lieu, les bordereaux de chiffrement et qu'il complétera ensuite par les messages d'anomalies. Si on veut lui éviter de manier les ciseaux et d'avoir a manipuler ensuite des morceaux de papier de toutes tailles il faudra faire en sorte qu'une page de listing ne contienne les messages que d'une seule US de niveau majeur.

.../...

7 - Regroupés par US, la encore afin de faciliter la tâche du gestionnaire pour qui il est plus commode de disposer de l'ensemble des messages d'une même US sur une seule page de listing même si les erreurs ont été détectées par des programmes différents.

8 - Standardisé enfin, il est souhaitable que d'une enquête à l'autre la présentation formelle des messages reste sensiblement la même.

Exemple : On pourra éditer les messages sous la forme ci-après

4 Identification de l'US est répété sur chaque ligne	1	Plage d'identification
	2	Textes des messages
	3	Image des données

La figure représente une page de listing.

1 - La plage d'identification situe le message dans le temps et dans l'espace. On y trouve donc la date d'édition, le n° d'ordre du tour de contrôle, les caractéristiques de la strate à laquelle appartient l'unité statistiques.

2 - La plage 2 donne la liste, en code et en clair des messages qui affectent l'US.

3 - La plage 3 contient l'information de base, sous forme d'image du questionnaire ou d'images de cartes, en totalité ou en partie selon le nombre et la nature des messages.

4 - La plage 4 répète sur chaque ligne l'identifiant de l'US.

.../...

n° d'ordre du passage — date — 2 lignes d'identification

```
C077011     ENTREPRISE  NO  3077011      SERVICE = NA     RP = 11     DEPART = 77             LOT NO 10        040673
C077011*****************************************************************************************************
```

liste des messages

```
C077011           MESSAGE  58  :  FORMES DES VENTES ERRONNEES
C077011           MESSAGE  60  :  LA SOMME DES INVESTISSEMENTS NE CORNESPOND PAS AU TOTAL INDIQUE
C077011                           INVEST =  0000000000
C077011                           SOMME INVESTIS = 0000025897
```

liste des cartes

```
C077011               01      01C077011 01747752293228    11 NA3574775229322841      1
C077011               02      02C077011 10       11                                      1  11090014212
C077011               03      03C077011              155000  1155000    5000  2000000
C077011               04      04C077011              20000                                              100
C077011               05      05C077011   40000    29000
C077011               06      06C077011
C077011               07      07C077011
C077011               08      08C077011212313   1550002514 1000000                                          1
C077011               10      10C077011               100  0
```

identifiant

III-2- Justification du programme d'édition des messages

Les caractéristiques 3 (messages triables), 5 (contenir l'information de base) et 7 (regrouper les messages d'une même US) conduisent à disjoindre la fonction d'édition des messages de la fonction de détection des erreurs.

En effet :

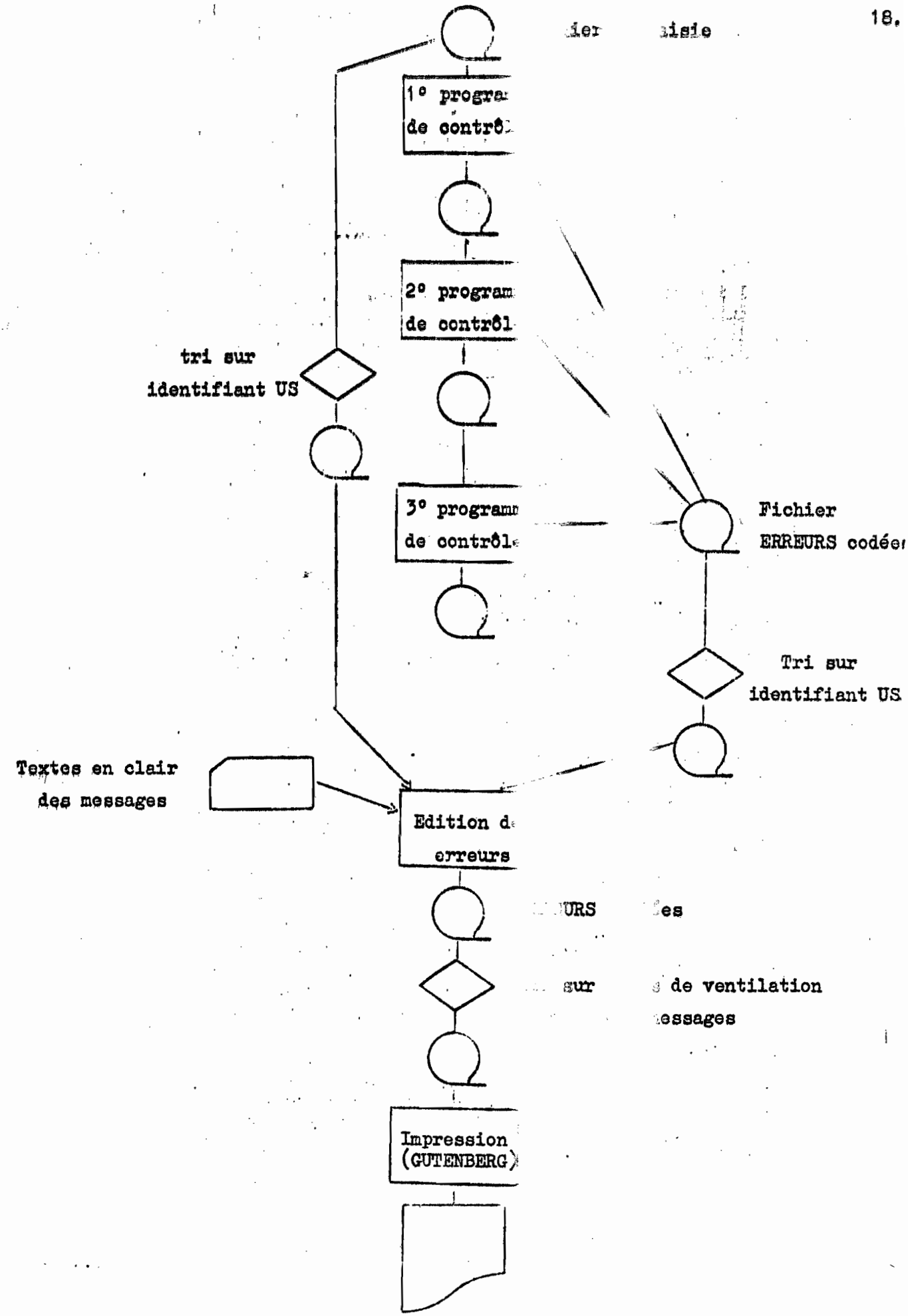
- L'ordre dans lequel doivent être éditées les US en anomalie n'est pas nécessairement celui dans lequel elles doivent être contrôlées. Dans le 1er cas il s'agit de faciliter la répartition des messages entre les ateliers et les équipes au sein des ateliers. Dans le 2ème il s'agira éventuellement de les classer selon des critères différents qui permettent par exemple de redresser une US en anomalie par les données de celle qui la précède.

- Au moment ou l'erreur ou anomalie est détectée l'information saisie peut n'être plus disponible parce qu'on aura redressé automatiquement certaines variables. Il faudra reprendre le fichier issu du carte à bande.

- L'ensemble des erreurs concernant une même US peuvent être détectées dans des programmes différents. Il faudra donc les regrouper.

On aboutira à une solution schématisée par l'organigramme ci-après, dans laquelle le ou les programmes de contrôle alimentent un fichier erreurs où elles sont stockées sous forme codée. Le fichier sera trié pour regrouper les erreurs par US puis pris en compte par le programme d'édition des erreurs qui prend comme 2ème entrée le fichier des données brutes de saisie pour produire un fichier des erreurs éditées conforme aux caractèristiques retenues.

.../...


1° progra
de contrô
2° program
de contrôl
tri sur
identifiant US
3° programm
de contrôl
Fichier
ERREURS codée
Tri sur
identifiant US
Textes en clair
des messages
Edition d
erreurs
Impression
(GUTENBERG)

.../.

III-3- Généralisation du programme d'édition des messages

Les tâches imparties au programme d'édition des messages sont bien définies et restent identiques d'une enquête à l'autre. Le programme est donc généralisable. Il en coûtera plus ou moins de temps en analyse et programmation selon qu'on cherchera à produire une édition très élaborée ou qu'on se contentera de quelque chose de relativement fruste comme dans l'exemple proposé ci-dessus. Il suffira de définir une structure standard du fichier "ERREURS CODEES" en confiant aux programmes de contrôle qui l'alimentent (et qui sont eux spécifiques) la définition complète des éléments qui doivent figurer dans le message de telle manière que le programme d'édition ait pour seule tâche la mise en page de ces éléments selon un standard invariant d'une enquête à l'autre.

Examinons plus en détail les problèmes à résoudre dans le cas d'une édition fruste d'abord, dans celui d'une édition plus élaborée ensuite.

III-3-1- Edition fruste

Il s'agit de produire des messages identiques à l'exemple ci-dessus.

Les données ont été saisies sur cartes perforées et on édite la totalité des données de l'US en anomalie, sous forme d'images de cartes, quelque soit le nombre et la nature des erreurs ou anomalies qui l'affectent.

La plage d'identification est limitée à une seule ligne.

Par erreur ou anomalie on édite deux ligne :

- La 1ère contient la partie dite fixe du message parce que son texte, qui décrit la nature de l'erreur ou anomalie, est invarian d'une US à l'autre. C'est celui dont le libellé devra être mis au point avec le plus grand soin afin qu'il soit compréhensible par tous et qui devra éventuellement être modifié s'il paraissait imprécis ou ambigu. Afin d'être aisément modifiables les textes fixes de messages devront se présenter comme des données externes au programme d'édition sous forme d'un petit fichier cartes ou disque qui ne devrait pas, dans la très grande majorité des cas, compter plus d'une centaine d'enregistrements, rare étant les enquêtes pour lesquelles il y a plus de 100 messages d'erreurs possibles.

.../...

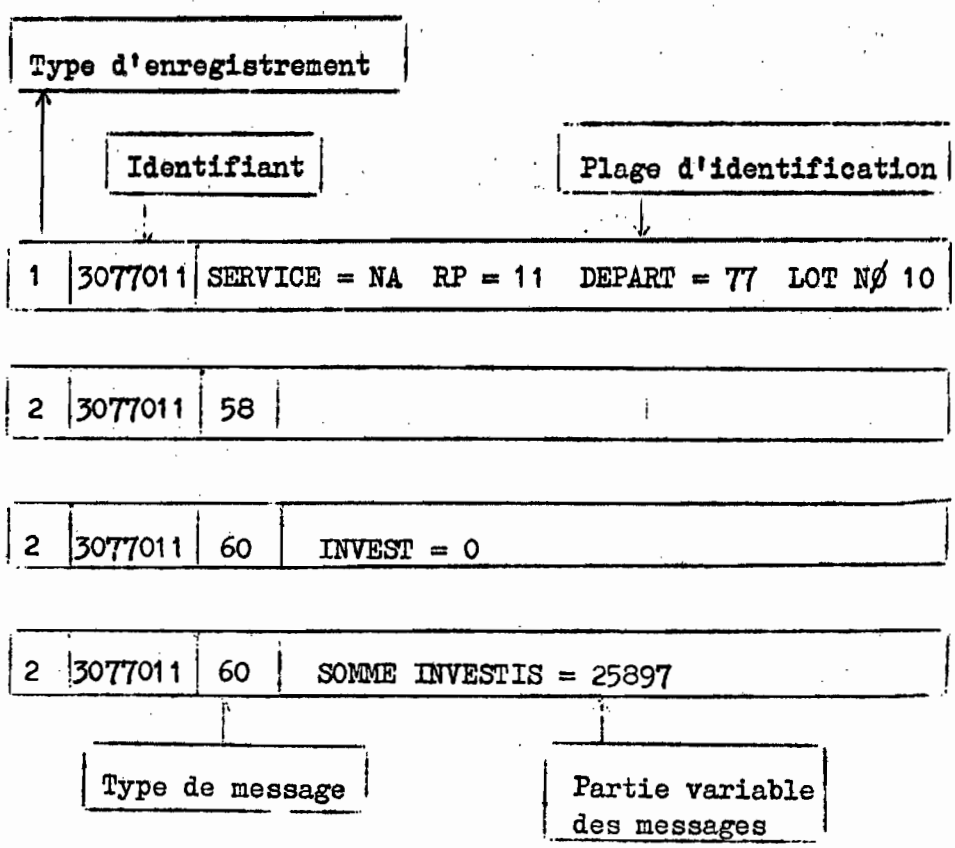
- La 2ème contient la partie dite variable. Il s'agit, dans tous les cas ou cela s'avère nécessaire, de rappeler les valeurs des variables incriminées par le contrôle. Il s'agit là d'indiquer les valeurs des variables au moment ou le contrôle est effectué, valeurs qui peuvent être différentes des valeurs origines telles qu'elles figurent dans les images de cartes dans la mesure ou un redressement automatique précédent le contrôle a pu les modifier (1).

Partant de ces spécifications on définira un fichier des erreurs codées contenant deux types d'enregistrements :

- le 1er fournira les données de la plage d'identification à l'exception de la date d'édition dont le calcul devra être confié au programme d'édition car contrôle et édition peuvent ne pas avoir lieu le même jour. L'enregistrement pourra être une image de la ligne éditée mise en forme par le programme de contrôle de telle manière que le programme d'édition ait pour seule tâche, sans paramètrage d'aucune sorte, de l'éditer en tête de page.

- le 2ème est relatif aux erreurs détectées. Il y a au moins un enregistrement par erreur dans lequel on trouve le code erreur et les éléments variables du message. Il pourra de même se présenter comme une image de la ligne éditée. Si les éléments sont en nombre tel qu'il faille plusieurs lignes pour les écrire tous on pourra avoir plusieurs enregistrements du 2ème type qui contiendront tous le même code erreur.

Exemple pour éditer la page de messages de l'exemple ci-dessus on créera les enregistrements suivants :



.../...

(1) ce qu'on devra s'efforcer d'éviter - voir paragraphe ci-après

Partant d'un tel fichier l'écriture du programme d'édition ne pose pas de problème sérieux dans la mesure où c'est le fichier qu'on a généralisé (en imposant une norme aux programmes de contrôle) et non pas le programme lui-même.

III-3-2- Edition élaborée

Il est assez malencontreux d'éditer les données sous forme d'images de cartes car c'est une forme peu lisible qui rend nécessaire l'utilisation d'un transparent sur lequel le dessin de carte a été reproduit ou de tout autre procédé permettant de repérer les variables dans les lignes éditées. De plus il n'est pas utile d'éditer l'ensemble de l'information relative à l'US quand le champ de l'erreur est bien délimité ; cela peut même se révèler gênant et coûteux quand cette information est très abondante.

On se donnera donc comme objectif complémentaire d'éditer l'information de base en partie ou en totalité dans une présentation qui rappelle celle du questionnaire. Il faudra alors associer aux différents types de messages des formats d'édition qui éditent dans une présentation soignée tout où partie de l'information. Ces formats devront être hiérarchisés ; certains d'entre eux peuvent contenir entièrement certains autres. Si deux formats présentant cette caractéristique sont à éditer pour une même US il faudra court-circuiter l'édition du moins pour éviter la redondance.

On pourra également associer à l'édition des messages celle d'un bordereau de correction "personnalisé" par la présence des données d'identification de l'US et la présence des cadres de corrections relatifs aux seules variables présumées erronées.

Le sous-système ESOPE du logiciel de dépouillement d'enquête LEDA offre un exemple intéressant de ce qu'on peut faire en ce domaine (voir annexe "Logiciels").

.../...

IV - TRI ET FUSION - LE TRI PAR GROUPES

Tri et fusion ne posent généralement pas de problèmes. On dispose dans tous les cas de produits-programmes, fournis par les constructeurs de matériels, qui donnent pleine satisfaction tant qu'il s'agit de trier ou fusionner des fichiers dans lesquels l'ensemble des critères de tri ou de fusion figurent dans chacun des enregistrements.

Il n'en ira pas de même avec des fichiers hiérarchisés qu'on voudra trier sur des critères qui figurent dans le seul enregistrement majeur. Pour reprendre l'exemple du paragraphe précédent on pourra vouloir trier le fichier des erreurs codées ou éditées sur les critères de la plage d'identification qui ne figurent que dans le 1er enregistrement de chaque US. On veut alors trier des groupes d'enregistrements sans modifier leur ordre à l'intérieur du groupe ou en le modifiant selon le cas. On parlera alors de tri par groupe.

IV-1- Principe du tri par groupes

Le tri comprend trois phases :

- la phase 1 a pour objectif de constituer en mémoire des séquences d'enregistrements triés et de les répartir sur les voies de tri. Le nombre d'enregistrements par séquence est fonction de la taille de ceux-ci et de l'espace mémoire à réserver à leur stockage,

- la phase 2 consiste à interclasser entre elles les séquences constituées lors de la 1ère phase pour en faire des séquences de plus en plus longues et de moins en moins nombreuses,

- la phase 3 est l'étape ultime de la phase 2, c'est à dire celle qui aboutit à la constitution d'une seule séquence qui est le fichier trié.

Pour faire un tri par groupe il suffira de :

- à l'entrée dans la phase 1, reporter les critères de tri qui ne figurent que dans l'enregistrement majeur en tête de tous les enregistrements du groupe de chaque US,

- laisser se dérouler la phase 2,

.../...

- à la sortie de la phase 3, reconstituer les enregistrements d'origines en effaçant les critères de tri.

IV-2- Solutions

Il existe deux solutions à ce problème :

- la première consiste à utiliser le tri standard du constructeur en lui adjoignant des exits de phase 1, pour le report des critères de tri, et de phase 3, pour leur suppression. Elle permet d'aboutir à une solution très générale mais requiert une bonne connaissance du fonctionnement du tri et du langage assembleur.

- la seconde consiste à écrire un programme en langage évolué (COBOL ou PL1) appelant le tri qui manipulera les enregistrements à l'entrée et à la sortie de celui-ci pour aboutir en même résultat. Elle est plus facile et plus rapide à mettre en oeuvre que la précédente mais plus difficile à généraliser.

Le sous-système TRISTAN de LEDA (cf annexe logiciels) constitue un exemple particulièrement évolué de tri par groupes.

V - PRODUCTION D'INDICATEURS SUR L'ETAT DU FICHIER

V-1- Les informations utiles

Il faut a tout moment pouvoir contrôler l'état du fichier et l'avancement du travail et pour cela connaître :

- le nombre d'US soumises à l'enquête,

- le nombre d'US dont les réponses ne sont pas encore parvenues,

- le nombre d'US dont les réponses sont parvenues mais n'ont pas encore été saisies,

- le nombre d'US saisies,

- parmi celles-ci le nombre de celles qui sont en erreur avec si possible une ventilation en deux ou trois postes selon la gravité des erreurs qui les effectent,

- le nombre d'US correctes ventilées selon la qualité de l'information entre correctes d'origine (sans redressement), partiellement redressées (une ou quelques variables), complétement redressées par extrapolation à partir d'une US voisine, par exemple,

- etc...

On devra pouvoir ventiler ces données selon quelques variables de référence tels que les critères de situation dans l'espace des messages d'anomalies et les grands critères de stratification de la population enquêtée. Ces variables devront être en nombre limité faute de quoi les tableaux qu'elles permettront de produire seront volumineux, difficilement exploitables et peu significatifs sauf aux niveaux les plus agrégés.

Il sera par ailleurs nécessaire, pour un contrôle plus affiné de la qualité du fichier et afin de permettre un jugement de valeur sur les questions posées et les réponses faites par les enquêtés d'abord, la validité des contrôles et redressements ensuite de disposer d'informations individuelles sur les variables et les résultats des contrôles.

.../...

Pour cela il faudra :

- au niveau de chaque US associer à chaque variable un code qualité indiquant par exemple si la variable est pure (sans redressement), redressée manuellement ou automatiquement,

- par contrôle, c'est à dire par message, connaître le nombre de ses occurrences.

Il importe de distinguer les comptages par variables des comptages par contrôles car les deux notions ne se recouvrent pas exactement. En effet un contrôle peut affecter plusieurs variables et à l'inverse une variable peut être affectée par plusieurs contrôles. Le comptage des variables apportera des élénents de jugement sur la qualité des données ; le comptage des contrôles permettra plutôt de vérifier leur pertinence.

Ces données seront ventilées selon les mêmes variables de références que les précédentes.

V-2- Utilisation des informations

Cet ensemble d'informations permettra de contrôler la qualité et l'exhaustivité du fichier et apportera des élénents non négligeables à une étude éventuelle de la refonte des questionnaires, des contrôles et des redressements. Il permettra également de gérer les rappels pour les enquêtes par voie postale, ce qui est, il est vrai, rarement voire jamais le cas des enquêtes dénographiques.

V-2-1- Contrôle de qualité

On pourra porter un jugement de valeur sur chacune des US soumises à l'enquête et sur chacune des variables observées. Ce qui permettra de juger de la valeur globale de l'enquête et d'en orienter le dépouillement statistique dans la mesure ou il apparait que les réponses à certaines questions ou de certaines catégories d'enquêtes sont de mauvaise qualité dans une proportion importante de cas. Les compteurs d'erreurs sont certes insuffisants pour bien juger de la qualité et ils devront être complétés par d'autres indicateurs, notamment des tableaux de contrôle (voir ci-après), mais ils permettent une 1ère approche.

.../...

V-2-2- Contrôle d'exhaustivité

Il s'agit de contrôler l'adéquation du fichier résultant au champ initial de l'enquête, mieux de faire en sorte quà tout moment on puisse savoir qu'elle est la proportion, relativement au champ de départ, d'enquêtes rentrées, saisies, correctes, en cours de correction. Les résultats de cet examen rapprochés de ceux du contrôle de qualité doivent permettre en particulier de savoir quand il sera possible de réaliser des exploitations provisoires du fichier, sur quels sous-ensembles du fichier elles peuvent porter, quel degré de finesse des résultats on peut obtenir, etc...

V-2-3- Correction des contrôles et redressements

Le fait qu'une erreur ou un redressement apparaisse très fréquemment peut provenir de ce que la question correspondante a été mal posée ou de ce que le contrôle qui lui est associé est erroné. Dans le 1er cas il est évidemment trop tard pour revenir en arrière, l'enquête étant faite. Dans le second, détecter précocement l'erreur d'analyse ou de programmation dès le contrôle du 1er lot par un examen du compteur associé permettra de corriger le programme incriminé avant de continuer le travail.

V-3- Solutions techniques

Il importe de disposer de certaines de ces informations au fur et à mesure des tours de contrôle ; ce sont celles relatives aux comptages des erreurs et redressements par types. Pour les autres, contrôle d'exhaustivité et de qualité, il faudra, à tout moment, être en mesure de répondre à la demande du statisticien.

Une 1ère solution consistera à prévoir dans le fichier permanent d'enquête une batterie de codes qualité relatifs aux US et aux variables et à associer à chaque programme de la chaîne d'apurement une batterie de compteurs relatifs aux erreurs détectées et aux redressements effectués. On obtiendra ainsi au fur et à mesure du passage des programmes, les comptages d'erreurs et redressements. Quant au contrôle d'exhaustivité et de qualité il suffira pour l'obtenir de tabuler le fichier permanent. Cette solution présente l'inconvénient de produire une information relativement dispersée en ce qui concerne les comptages d'erreurs et d'anomalies (une série de compteurs par programme) et d'être assez lourde en ce qui concerne le contrôle de qualité et d'exhaustivité puisqu'il faudra pour l'obtenir tabuler l'ensemble du fichier permanent. De plus elle ne se prête à aucune généralisation.

.../...

Une seconde solution consistera à rassembler toutes les données de gestion dans un fichier spécialisé qu'on nommera fichier DIRECTEUR. Il contient au départ pour seule information la liste des identifiants des US soumises à l'enquête. On y chargera au fur et à mesure de la prise en compte et du contrôle des US, toutes les informations énumérées ci-dessus dans un enregistrement qui contiendra par exemple les zones suivantes :

- Identifiant d'US,

- Zone "variables de gestion" qui sont les critères de situation et les grands critères de stratification,

- Code état (l'US à répondu, a été saisie, etc...

- Code qualité de l'US,

- Zone codes qualité par variables.

Dès l'instant qu'on dispose d'un fichier dont la structure est standardisée il sera relativement aisé, même si la longueur des zones varie d'une enquête à l'autre, de réaliser un programme d'exploitation de ce fichier qui serve pour toutes les enquêtes.

Une variante consistera à confier au programme d'édition des ERREURS, l'édition des comptages d'erreurs et de redressements.

VI - CONTROLES STATISTIQUES - HISTOGRAMMES

L'objet des contrôles statistiques est de permettre un jugement global sur la qualité du fichier avant, pendant ou après l'apurement :

- le contrôle avant apurement permettra de détecter certains points faibles de l'enquête et par conséquent d'orienter l'effort de correction,

- le contrôle en cours, où du moins en début d'apurement permettra de détecter d'éventuelles erreurs sur les contrôles et redressements individuels,

- le contrôle après apurement en apportant des indications sur la qualité finale du fichier permettra d'en orienter judicieusement l'exploitation.

Les indicateurs décrits au paragraphe précédent permettent de réaliser une part appréciable des contrôles statistiques en donnant tous renseignements utiles sur l'avancement et la qualité du travail (on parlera alors plus volontiers de contrôle de gestion). Par contre ils apportent peu d'informations sur la structure statistique du fichier. Pour cela il faudra avoir recours aux tableaux de contrôle.

Les tableaux de contrôle sont des tableaux au sens habituel du terme. Du point de vue informatique rien ne les distingue des tableaux d'analyse puisqu'ils seront produits par les mêmes procédures. Toutefois les logiciels de tabulation exigent habituellement que les fichiers qu'on leur donne à traiter soient corrigés, structurés, codifiés. Le langage de description des tableaux peut atteindre une certaine complexité, d'autant plus grande que les possibilités du logiciel sont plus étendues. Ceci à pour conséquence qu'on ne peut produire des tableaux de contrôle qu'après que le fichier ait subi un minimum de manipulations, ce qui prend du temps. D'où l'intérêt d'un logiciel qui permette de produire, à un très faible coût de paramètrage, des tableaux simples à partir des données brutes saisies. Si la mise en oeuvre en est asses simple pour permettre la production de quelques tableautins dans les quelques heures qui suivent la mise à disposition des données sur support magnétique on pourra détecter les erreurs systématiques d'enquête, chiffrement ou saisie les plus grossières assez tôt pour pouvoir prendre les mesures correctrices.

De tels logiciels seront appelés logiciels d'HISTOGRAMMES car ils ne produiront habituellement que des tableaux à une seule dimension.

.../...

Exemple très simple

Le programme traite des enregistrements en format carte perforée. Pour chaque type de carte et colonne par colonne il donne la ventilation des valeurs qui y sont inscrites. Par type de carte il édite un tableau identique à celui de la figure ci-jointe.

Un tel programme d'histogrammes à pour principal avantage d'être d'une extrême simplicité tant en ce qui concerne sa réalisation que sa mise en oeuvre.

Par contre ses possibilités sont extrêmement limitées puisque, dans le mode présentation proposé, on ne pourra pratiquement analyser la ventilation de la population que selon des codes à un seul chiffre ce qui peut se révéler suffisant pour les enquêtes démographiques dont les codes comportent rarement un grand nombre de postes.

On peut imaginer (il en existe d'ailleurs) des programmes d'histogrammes qui :

- traitent des enregistrements de longueur quelconque,

- calculent et éditent des fréquences par quantiles pour les variables quantitatives,

- ventilent selon des codes comportant un nombre élevé de postes,

-permettent de sélectionner des sous-populations qui seules seront analysées,

- etc...

On en trouvera un exemple dans LEDA avec le sous-système ORESTE (voir annexe LOGICIELS).

.../...

CARTE DE TYPE 01

	VALEURS											
	0	1	2	3	4	5	6	7	8	9	' '	Autre
COLONNE 3	3	10	80	10	5	0	0	1	0	0	8	4
COLONNE 4	0	0	0	0	0	0	0	0	0	0	120	0
COLONNE 5	0	47	48	0	1	0	0	1	0	0	2	1
COLONNE 80	0	0	0	0	0	0	0	0	0	0	120	0

NOMBRE TOTAL DE CARTES 01 = 120

Exemple d'histogramme simple

- Le type de carte est en colonnes 1 et 2 dont il n'y a pas lieu de ventiler le contenu.
- La colonne 3 contient l'état matrimonial dont les valeurs correctes sont : 1 = célibataire, 2 = marié, 3 = veuf, 4 = divorcé, la ventilation est à priori incorrecte.
- La colonne 4 est vide.
- La colonne 5 contient le sexe dont les valeurs correctes sont : 1 = homme, 2= femme, la ventilation aux valeurs erronées près parait correcte.

VII - CONTROLES ET REDRESSEMENTS INDIVIDUELS

Les contrôles statistiques renseignent sur la qualité globale du fichier. Ils permettent également de détecter une partie au moins des erreurs systématiques qui résultent d'une interprétation erronée, des instructions ou d'un défaut de ces instructions, à quelque stade de l'enquête que ce soit (sur le terrain, au chiffrement, à la saisie,...).

Les contrôles et redressements individuels ont pour objectif de détecter et corriger, unité statistique par unité statistique, des anomalies qui à priori sont aléatoires. Il se peut qu'à cette occasion on détecte également des erreurs systématiques qui auraient échappé au contrôle statistique. En tout état de cause la procédure de redressement de l'erreur systématique restera, quelque soit la façon dont on l'a détectée, différente de la procédure de redressement de l'erreur aléatoire. Pour la seconde, tout naturellement, la correction est individualisée, soit par retour au questionnaire, soit par une procédure automatique qui recherchera la valeur vraie de la variable erronée dans le contexte de l'unité statistique ou par référence à une unité statistique de même structure. Pour la première on cherchera plutôt, et ce de manière automatique, à dérouler à l'envers le processus qui a conduit à la valeur fausse pour retrouver la valeur vraie. Lorsque cela se révèlera impossible on devra se résoudre à n'utiliser qu'avec la plus grande prudence la variable incriminée (voire à ne pas l'utiliser du tout) car les procédures de redressement automatique appliquées aux erreurs aléatoires ne peuvent ici donner que des résultats médiocres ; et un redressement manuel systématique sera généralement trop coûteux pour être envisagé.

Les contrôles et redressements individuels constituent le noeud et la principale raison d'être de la chaîne d'apurement. Ils en sont aussi la plus grande difficulté. Leur réalisation représente la plus grosse part du travail d'analyse et programmation car la généralisation et le recours à des programmes ou procédures standardisées est le plus souvent impossible chaque enquête constituant un cas bien particulier. De plus se pose le problème de la hiérarchisation dont on a vu dans le chapitre 2 qu'il n'avait pas de solution générale. Nous allons dans la suite de ce paragraphe dégager des règles d'organisation et de traitements simples qui devraient permettre d'éviter les principaux écueils.

.../...

VII-1- Contrôles à redressement manuel et contrôles à redressement automatique

Une erreur peut être corrigées automatiquement ou manuellement. Dans le 1er cas la correction est faite par le programme qui a détectée l'erreur en fonction d'un contexte plus ou moins étendu. Dans le second le programme émet un message d'erreur remis au gestionnaire qui rédigera un bordereau de correction qui viendra après saisie corriger la donnée erronée. Pour les contrôles à redressement automatique il suffira d'un seul passage du programme (supposé correct) pour que les données incriminées soient corrigées. Pour les contrôles à redressement manuel par contre il faudra, une fois la donnée correctrice introduite, contrôler à nouveau l'US car le correcteur humain n'est pas, contrairement au programme, d'une complète fiabilité. Il y aura donc itération jusqu'à ce que le programme de contrôle ne détecte plus aucune erreur.

Dans certaines enquêtes tous les redressements sont automatiques, dans d'autres ils sont manuels, dans d'autres encore on mêle redressements automatiques et manuels.

1er cas - tous les redressements sont automatiques

Le principal problème sera celui de la hiérarchisation des contrôles. Les variables erronées vont être corrigées en fonction d'un contexte externe (exemple : le hot-deck) ou interne parce qu'il existe une corrélation entre la variable erronée et une ou plusieurs autres variables de l'US. Il faudra alors s'assurer que la ou les variables redressantes sont elles-même correctes ou ont une probabilité très forte de l'être, faute de quoi on peut arriver à remplacer une valeur primitivement correcte par une valeur erronée. C'est avant tout au statisticien qu'il appartient de résoudre ce problème. L'informaticien devra toutefois constamment s'en préoccuper dans la phase d'analyse qui consiste à organiser les contrôles dans les programmes car cette opération d'agencement des traitements est la meilleure occasion de détecter les oublis ou erreurs commis lors de la définition des contrôles.

.../...

2ème cas - Tous les redressements sont manuels

Le problème de la hiérarchisation des contrôles se pose à priori avec moins d'acuité que dans le cas précédent. Le correcteur étant un être pensant, on peut en effet espérer qu'il ne modifiera pas aveuglèment une variable manifestement correcte pour laquelle un message a été émis à tort. Il n'empêche que si on ne prend aucune précaution les gestionnaires vont se trouver noyés sous une avalanche de messages bizarres qui les conduira à douter de la santé mentale des informaticiens ! Les motifs de friction entre les uns et les autres sont par ailleurs assez nombreux pour qu'on les multiplie à plaisir.

La hiérarchisation des contrôles devra donc rester une préoccupation majeure. On pourra par ailleurs clarifier les relations en :

- informant correctement les gestionnaires de l'économie générale du système de contrôle et des limites intellectuelles de l'outil informatique,

- prévoyant dans les programmes, une fois tous les contrôles effectués et avant de déclencher l'emission des messages, un examen de l'ensemble des erreurs détectées. Certaines configurations d'anomalies révèleront en effet de façon quasi-certaine la présence de fausses erreurs. Si, par exemple, une variable pivot sert au contrôle de plusieurs autres et si celles-ci sont toutes en erreur cela induit presque sûrement que c'est la variable pivot qui est erronée. On émettra alors un message du type suivant :

"LA VARIABLE X EST PROBABLEMENT ERRONEE ; LES MESSAGES RELATIFS AUX VARIABLES Y1, Y2 ... ONT PEUT ETRE ETE EMIS A TORT - PROCEDEZ A UN REEXAMEN COMPLET DU QUESTIONNAIRE.

3ème cas - Redressements manuels et redressements automatiques

A priori on additionne les problèmes des deux cas précédents.

Toutefois il convient de se souvenir que si pour les redressements manuels il y a nécessairement itération du programme de contrôle, il suffit pour les redressements automatiques d'un seul passage.

On admettra par ailleurs que lorsqu'on mêle les deux modes de redressement le choix de l'un ou l'autre sera, variable par variable, guidé par les critères suivants :

1 - le redressement manuel sera appliqué aux variables clés les plus importantes qui définissent, pour l'essentiel, la structure de l'US,

.../...

2 - le redressement automatique est par contre appliqué aux variables secondaires qu'on peut aisément estimer à partir des variables clés dès l'instant que celles-ci sont correctes,

3 - pour une variable clé on pourra malgré tout choisir le redressement automatique dès lors que la divergence par rapport à la valeur supposée vraie est faible. Ce sera en particulier le cas des contrôles de balance pour lesquels un faible écart en valeur absolue ou relative pourra sans dommage être corrigé automatiquement.

Dès lors il devient possible et même souhaitable de répartir contrôles à redressements manuels et contrôles à redressements automatiques dans des programmes distincts. Le premier chargé des contrôles à redressement manuel présentera les caractéristiques simplifiées du 2ème cas. Il itérera jusqu'à résorbtion complète des erreurs. Le 2ème qui marquera la phase ultime de l'apurement, sera passé une seule fois avant codification et tabulation. On aura pas à s'y soucier de hiérarchiser contrôles et redressements puisque ces derniers se feront uniquement à partir des variables clés qui à ce stade sont toutes correctes.

On pourra d'ailleurs mêlés les contrôles et redressements automatiques et le polissage (voir ci-après) dans le même programme puisque [illegible] opérations interviennent au même point du processus d'apurement.

Adopter cette solution, suppose que le choix entre variables à redresser manuellement et variables à redresser automatiquement soit parfaitement judicieux. Il est à peu près impossible de définir des règles, chaque enquête ou type d'enquête constituant un cas particulier. En tout état de cause une réflexion approfondie sur ce problème amènera à s'interroger sur la pertinence des contrôles et dégage une méthode rationnelle d'élaboration et de clarification des contrôles à effectuer ; la fiabilité du sous système d'apurement s'en trouvera renforcée dès le stade de l'analyse fonctionnelle, l'analyse organique et l'écriture des programmes s'en trouveront facilité ainsi que la compréhension des messages par les gestionnaires.

.../...

VII-2- Les principaux types de contrôles individuels

Le paragraphe précédent propose une première partition entre contrôles à redressement manuel et contrôles à redressement automatique. On peut également les ventiler selon la nature du concept contrôlé. De ce point de vue nous les répartirons en trois grands groupes qui sont :

- les contrôles de la mise à jour,
- les contrôles de structure,
- les contrôles de validité et cohérence des données.

VII-2-1- Les contrôles de la mise à jour

Nous y reviendrons plus loin dans le paragraphe sur la mise à jour. Disons seulement ici qu'ils n'auront, en principe, à intervenir que lorsque l'enquête utilise les redressements manuels. La mise à jour aura alors pour principale fonction d'introduire les modifications de variables proposées par les gestionnaires après qu'elles aient été déclarées erronées par les programmes de contrôle. La correction des erreurs de mise à jour se fera donc manuellement.

VII-2-2- Les contrôles de structure

L'objet des contrôles de structure est de s'assurer que pour toute US du plus haut niveau, on dispose bien de toutes les pièces obligatoires du dossier qui la décrivent. En d'autres termes, cela reviendra à vérifier la validité de l'arborescence qui décrit les différents types d'enregistrements et leurs enchaînements.

Ils n'auront bien sur à intervenir que dans la mesure où l'enquête étudie plusieurs types d'unités statistiques liées entre elles par des liens hiérarchiques.

Ils précèderont logiquement les contrôles de validité et cohérence des données car on ne peut guère envisager ces derniers si les données sont incomplètes.

.../...

Enfin ils feront partiedes contrôles à redressement manuel dans la mesure ou ceux-ci sont envisagés. C'est d'ailleurs le seul cas que nous considérerons dans la suite de ce paragraphe. Si l'enquête n'utilise que des redressements automatiques la méthode exposée reste toutefois valable à ceci près que l'émission des messages d'erreurs devra être remplacée par une procédure d'estimation des données manquantes ou de rejet de l'unité statistique incriminée qui sera spécifique à chaque enquête.

VII-2-2-1- Délimitation du champ

Le contrôle de structure présente un aspect formel et un aspect logique.

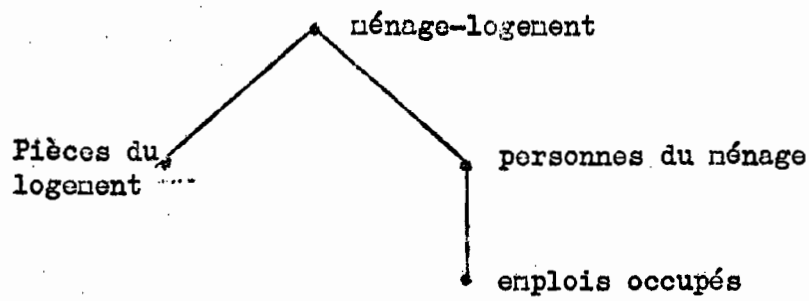
Aspect formel (cf chapitre sur les fichiers)

Il consistera seulement à vérifier qu'il ne manque pas de maillons dans l'enchaînement obligatoire des types d'enregistrements qu'implique toute hiérarchisation d'un fichier.

Considérons une enquête auprès des ménages dans laquelle on a défini les sous-questionnaires suivants :

- ménage - logement,
- pièces du logement,
- personnes du ménage,
- emplois occupés par les personnes du ménage au cours de l'année précédent l'enquête.

On aboutira logiquement à une structure du dossier, donc du fichier, du type suivant :



.../...

Le contrôle formel consistera dans ce cas à s'assurer que tout sous-questionnaire emploi occupé est bien rattaché à un sous-questionnaire personne et que tout sous-questionnaire personne ou pièce de logement est bien rattaché à un sous-questionnaire ménage-logement.

A l'inverse, le cas d'un sous-questionnaire ménage-logement non suivi de sous-questionnaires pièces ou (et) personne sera considéré comme normal (le logement peut être vide, on peut ne pas disposer de l'information relative aux pièces du logement), de même que celui du sous-questionnaire personne non suivi de sous-questionnaires emplois.

Aspect logique

Il consistera à s'assurer de la présence effective de sous-questionnaires mineurs dont la présence est impliquée par des variables du majeur auxquels ils sont rattachés (si le sous-questionnaire ménage-logement indique qu'il y a 6 personnes dans le ménage, on devra s'attendre à trouver 6 sous-questionnaires personne à la suite ; si le sous-questionnaire personne contient le total des rémunérations perçues et le sous-questionnaire emploi la rémunération de l'emploi et qu'on constate que la sommation des rémunérations par emploi est très inférieure au total du niveau supérieur, on pourra présumer qu'il manque des sous-questionnaires emploi).

Nous exclurons ce 2ème aspect des contrôles de structure stricte-sensu. On peut en effet considérer que, dès l'instant qu'ils font intervenir les données elles-mêmes, on entre dans le champ des contrôles de cohérence. De plus, rien n'indique que l'anomalie constatée est bien structurelle ; elle peut aussi bien provenir de la variable de contrôle. Enfin, comme on va le voir, on peut appliquer au contrôle de structure formel, quelle que soit l'arborescence du fichier, un algorythme assez général que la prise en compte de l'aspect logique perturberait fortement.

VII-2-2-2- Méthodologie du contrôle de structure

Le contrôle de structure peut devenir extrêmement délicat et se heurter à de grandes difficultés de mise au point pour peu que l'arborescence du fichier contrôle présente quelque complexité. C'est pourquoi il est intéressant de chercher à définir une méthode d'approche du problème, débouchant sur un algorithme aussi général que possible, qui évitera à chaque nouvelle enquête les erreurs et tatonnements générateurs de retards et de déboires de tous ordres.

Celle que nous allons exposer n'est pas unique. Elle a le mérite d'être assez simple et de bien illustrer ce qu'on peut faire en ce domaine. Nous l'éclairerons en l'appliquant au fichier "ménage-logement", cité ci-dessus qui est assez complexe. Afin de bien préciser l'exemple, nous supposerons que les données sont saisies sur cartes perforées. Enfin, les cartes devront être correctement triées avant d'entrer dans l'algorithme de contrôle de structure.

.../...

VII-2-2-2-1- Structure du fichier

- Les données "ménage-logement" occupent deux cartes, de types 1 et 2, qui sont donc les seules toujours obligatoires. L'identifiant ménage-logement qui figure dans tous les types de cartes est nommé ID1,

- les données "pièce du logement" occupent une carte de type 3 ; l'identifiant "pièce" est nommé ID3 et vient s'ajouter dans la carte de type 3 à ID1 ; la carte de type 3 est facultative multiple par rapport au groupe des cartes 1 et 2,,

- les données "individu" occupent deux cartes de type 4 et 5. L'identifiant individu est nommé ID4 et vient s'ajouter dans les cartes 4 et 5 à ID1. Le groupe 4-5 est facultatif multiple par rapport au groupe 1-2 ; la présence d'une carte 4 rend obligatoire la présence d'une carte 5 de même identifiant ID4 et inversement,

- les données "emploi" occupent une carte de type 6. L'identifiant emploi est nommé ID6. La carte 6 est facultative multiple par rapport au groupe 4-5 ; l'identifiant ID6 vient s'ajouter à ID1 et ID4.

La structure du fichier peut au total se résumer comme suit en ce qui concerne la séquence des types de cartes :

$$1,2,\ [3,\ [3]\ ,\ldots]\ ,\ [4,5,\ [6,\ [6]\ ,\ldots\ ]\ ,\ [4,5,\ [6,\ [6]\ ,\ldots]\ ,\ldots$$

VII-2-2-2-2- Tri du fichier

En préalable au contrôle de structure, les cartes doivent être triées dans l'ordre des identifiants soit selon les critères suivants :

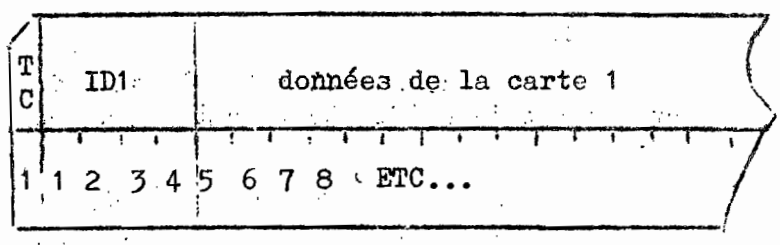
ID1 x ID3 x ID4 x ID6 x TC (type de carte)

Ce tri pourra poser problème si dans une carte décrivant une US de niveau supérieur la zone, qui dans la carte décrivant une US de niveau inférieur contient l'identifiant le 2ème rang, est occupée par des codes qui viendront parasiter le tri.

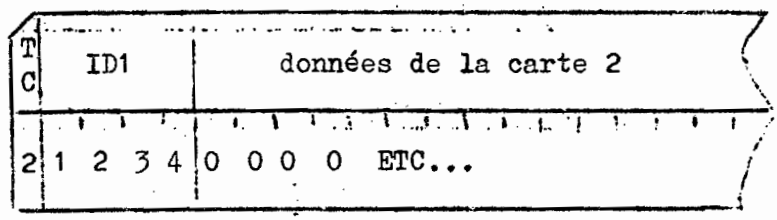
.../...

Exemple : supposons les cartes 1, 2 et 3 avec les dessins et les valeurs suivantes :

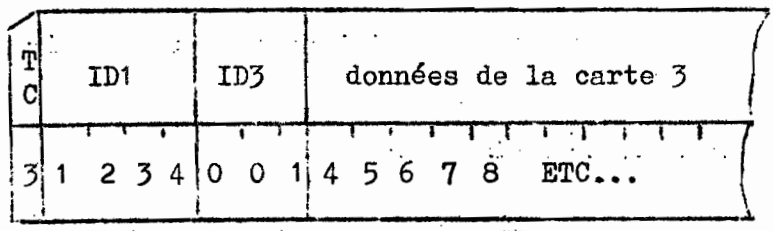
CARTE 1



CARTE 2



CARTE 3



Si on trie ces 3 cartes selon les critères indiqués ci-dessus la carte 1 dont les 3 premiers caractères de données seront considérés comme un ID3, va se trouver placer derrière la carte 3 et on obtiendra la séquence (2, 3, 1) qui est incorrecte.

.../...

Deux solutions peuvent pallier cet inconvénient :

1) Laisser vide dans les cartes de rang supérieur les zones qui correspondent aux identifiants des cartes de rang inférieur : on risque si la structure est assez complexe donc s'il y a beaucoup d'identifiants de perdre beaucoup de place sur les cartes,

2) Modifier les dessins de cartes avant le tri en "sortant" les identifiants pour les groupes tous en tête, ce qui donnera :

	ID1	ID3	ID4	ID6	TC		
carte 1	1 1 1 1	-	-	-	1	données	
carte 2	1 1 1 1	-	-	-	2	données	
carte 3	1 1 1 1	0 0 1	-	-	3	données	

•
•
•
etc

Cette modification peut se faire par un exit de phase 1 au tri.

Remarque importante :

Si on veut que les cartes 3 suivent immédiatement les cartes 1 et 2, il faudra en fait classer les critères de tri dans l'ordre.

ID1 x ID4 x ID6 x ID3 x TC

Les cartes 3 dont les ID4 et ID6 sont nuls se trouveront ainsi bien placées devant les cartes 4, 5 et 6 (1).

---

(1) Ceci n'est d'ailleurs pas impératif. Les US "pièce" et "individu" étant des descendants de même rang du "ménage-logement" leur ordre importe peu car elles n'ont pas de préséance logique l'une par rapport à l'autre. Il faudra, si le tri est tel que des groupes (4, 5, 6,...) peuvent précéder des cartes 3, admettre comme correctes, dans le tableau de la page suivante, les séquences 53 et 63.

.../...

VII-2-2-2-3- Le contrôle

Le contrôle s'opère au niveau de l'unité statistique de plus haut niveau, soit le ménage-logement dans notre exemple.

Toute US du plus haut niveau comportant une erreur de structure sera rejetée, même si l'erreur concerne seulement un de ses sous-questionnaires.

Le nombre de configurations qui peuvent représenter une structure valide est dans notre exemple, et à fortiori dans une structure plus complexe, très élevé. Vouloir en dresser la liste exhaustive serait long, fastidieux et comporterait des risques nombreux d'erreurs et d'omissions.

C'est pourquoi on examinera les cartes d'une US majeure, c'est à dire de même identifiant ID1 dans notre exemple, deux à deux et on examinera à chaque fois si la séquence de deux cartes observée est correcte ou non. Ainsi :

- la séquence carte 1 - carte 2 peut être jugée correcte puisqu'on a retenu un ensemble de cartes ayant même ID1,

- la séquence carte 2 - carte 3 est également correcte,

- une séquence de deux cartes 3 successives sera correcte si elles ont des ID3 différents ; si ID3 a la même valeur pour les deux, il s'agira d'un double,

- une séquence carte 4 - carte 5 sera correcte si elles ont même ID4 ; sinon, cela voudra dire qu'il manque la carte 5 du 1er individu pour lequel on a la carte 4 et la carte 4 d'un deuxième pour lequel on a la carte 5,

- etc...

On examinera ainsi la séquence 1ère carte - 2ème carte, puis la séquence 2ème carte - 3ème carte, et ainsi de suite. Il faudra également s'assurer que le groupe commence bien par une carte 1 et se termine bien par une carte 2, 3, 5 ou 6.

.../...

L'ensemble des cas possibles peut être recensé dans le tableau ci-dessous :

1ère carte	2ème carte													
	1		2		3		4		5		6		rien	
Rien	01	C	02	E	03	E	04	E	05	E	06	E	07	E
1	11	E	12	C	13	E	14	E	5	E	16	E	17	E
2	21	E	22	E	23	C	24	C	25	E	26	E	27	C
3	31	E	32	E	33	ID3	34	C	35	E	36	E	37	C
4	41	E	42	E	43	E	44	ID4	45	ID4	46	E	47	E
5	51	E	52	E	53	E	54	C	55	ID4	56	ID4	57	C
6	61	E	62	E	63	E	64	C	65	E	66	ID4 ID6	67	C

dans lequel on a noté :

- <u>en 1ère sous-case</u> :

La valeur de la séquence de types de carte observée en notant 0 l'absence de la 1ère carte (on examine la 1ère carte du lot) et 7 l'absence de la dernière carte (on examine la dernière carte du lot).

.../...

- en 2ème sous-case

- "C" si la séquence est correcte,

- "E" si la séquence est erronée,

- "IDx" (x parmi 3, 4, 6) s'il est nécessaire de comparer les valeurs d'ID3, ID4 ou ID6 dans les 2 cartes de la séquence pour compléter le contrôle.

Le lecteur pourra vérifier la validité du tableau.

Ce tableau est insuffisant pour déterminer les contrôles. En effet :

1 - On peut trouver des cartes dont le type est erroné (différent de 1 à 6). Une telle carte sera rejetée avec le message "TYPE DE CARTE ERRONE" et sera négligée dans la suite des contrôles. Si on a la séquence "1, 2, 9, 3", "2, 9" provoquera le message ci-dessus et on passera ensuite à l'examen de "2, 3".

2 - En cas d'erreur, il s'agit de déterminer la nature de l'erreur (carte manquante, carte en double, carte déclassée, tous types de cartes absents dans le cas ou tous les types de cartes sont erronés) et les types de cartes incriminés.

3 - Dans le cas où il faut comparer un identifiant, il faudra déterminer ce qui résulte de cette comparaison.

C'est pourquoi on développera le contenu du tableau dans la table ci-après dans laquelle :

- la colonne 1 donne le n° de ligne de la table,

- la colonne 2 donne la valeur de la séquence examinée,

- la colonne 3 donne un code retour (CR) qui prend les valeurs :

  . 1 si la séquence est correcte,

  . 2 si la séquence est erronée,

  . 3 s'il faut affiner le contrôle par la comparaison d'un identifiant,

.../...

- la colonne 4 donne le code erreur (CE) qui prend les valeurs :

. 1 pour une carte manquante,

. 2 pour une carte en double,

. 3 pour une carte déclassée,

. 4 si tous les types de cartes sont absents,

CE n'est renseigné que lorsque CR = 2.

- la colonne 5 donne les types de cartes en erreur (TCE). TCE n'est renseigné que lorsque CE est égale à 1, 2 ou 3.

- la colonne 6 (ID) donne le nom de l'identifiant à comparer lorsque CR = 3,

- la colonne 7 (RETE) donne l'adresse de la table où se trouve décrite l'action à entreprendre lorsque l'identifiant comparé à même valeur dans les deux cartes (CR = 3 et ID renseigné),

- la colonne 8 (RETD) donne l'adresse de la table où se trouve décrite l'action à entreprendre lorsque l'identifiant comparé à des valeurs différentes dans les deux cartes (CR = 3 et ID renseigné).

T A B L E D E C O N T R O L E

N° de ligne (1)	séquence contrôlée (2)	CR (3)	CE (4)	TCE (5)	ID (6)	RETE (7)	RETD (8)
1	01	1					
2	02	2	1	1			
3	03	2	1	1, 2			
4	04	2	1	1			
5	05	2	1	1, 2, 4			
6	06	2	1	1, 2, 4, 5			
7	07	2	4				
8	11	2	2	1			
9	12	1					

.../...

N° de ligne (1)	séquence contrôlée (2)	CR (3)	CE (4)	TCE (5)	ID (6)	RETE (7)	RETD (8)
10	13	2	1	2			
11	14	2	1	2			
12	15	2	1	2, 4			
13	16	2	1	2, 4, 5			
14	17	2	1	2			
15	21	2	3	1			
16	22	2	2	2			
17	23	1					
18	24	1					
19	25	2	1	4			
20	26	2	1	4, 5			
21	27	1					
22	31	2	3	1			
23	32	2	3	2			
24	33	3			ID3	50	51
25	34	1					
26	35	2	1	4			
27	36	2	1	4, 5			
28	37	1					
29	41	2	3	1			
30	42	2	3	2			
31	43	2	3	3			
32	44	3			ID4	52	53
33	45	3			ID4	56	55
34	46	2	1	5			
35	47	2	1	5			
36	51	2	3	1			
37	52	2	3	2			
38	53	2	3	3			
39	54	1					

.../...

N° de ligne (1)	séquence contrôlée (2)	CR (3)	CE (4)	TCE (5)	ID (6)	RETE (7)	RETD (8)
40	55	3			ID4	56	57
41	56	3			ID4	58	59
42	57	1					
43	61	2	3	1			
44	62	2	3	2			
45	63	2	3	3			
46	64	1					
47	65	2	1	4			
48	66	3			ID4	60	61
49	67	1					
50	33	2	2	3			
51	33	1					
52	44	2	2	4			
53	44	1	1	5			
54	45	1					
55	45	2	1	5, 4			
56	55	2	2	5			
57	55	1	1	4			
58	56	1					
59	56	2	1	4, 5			
60	66	3			ID6	62	63
61	66	2	1	4, 5			
62	66	2	2	6			
63	66	1					

.../...

Là encore, le lecteur est invité à contrôler la validité de la table.

Examinons en détail le cas le plus complexe qu'elle contienne, celui où on a une séquence de deux cartes 6 :

- à la ligne 48 on indique qu'il faut comparer ID4 et qu'il faut aller à la ligne 60 en cas d'égalité et à la ligne 61 si non,

- à la ligne 60 (ID4 =, ce qui veut dire que la 2ème carte 6 est bien rattachée au même groupe 4-5 que la première), on indique qu'il faut comparer ID6 pour s'assurer qu'il ne s'agit pas d'un double et on renvoie à la ligne 62 en cas d'égalité et à la ligne 63 dans le cas contraire,

  . à la ligne 62 (ID6 =) on indique que l'erreur est détectée (CR = 2) et qu'il s'agit d'une carte 6 en double (CE = 2, TCE = 6),

  . à la ligne 63 (ID6 ≠) on constate que la séquence est correcte (CR = 1),

- à la ligne 61 enfin, on constate que deux cartes 6 successives ont des ID4 différents, ce qui signifie qu'il manque les cartes 4 et 5 auxquelles se rattache la deuxième (CE = 1, TCE = 4, 5).

Généralisation

Elle est évidente. Une telle table s'adapte à n'importe quelle structure de fichier aussi complexe soit-elle.

Le principe en est simple mais il faudra veiller à la renseigner avec soin car le risque d'erreur n'est pas inexistant.

Elle peut être aisément stockée en mémoire centrale vu son faible encombrement : dans notre exemple, elle compte approximativement 950 caractères pour une arborescence relativement complexe ; il ne devrait donc pas y avoir, dans la grande majorité des cas, de problème de place mémoire.

Quant à l'algorithme de consultation dont nous donnons un exemple pas complètement détaillé ci-après, il sera toujours simple, standard et indépendant de la structure elle-même. Il pourra donc être réutilisé au prix de faibles modifications (ou tel quel, si on a songé à le paramétrer pour le généraliser) d'une enquête à l'autre.

.../...

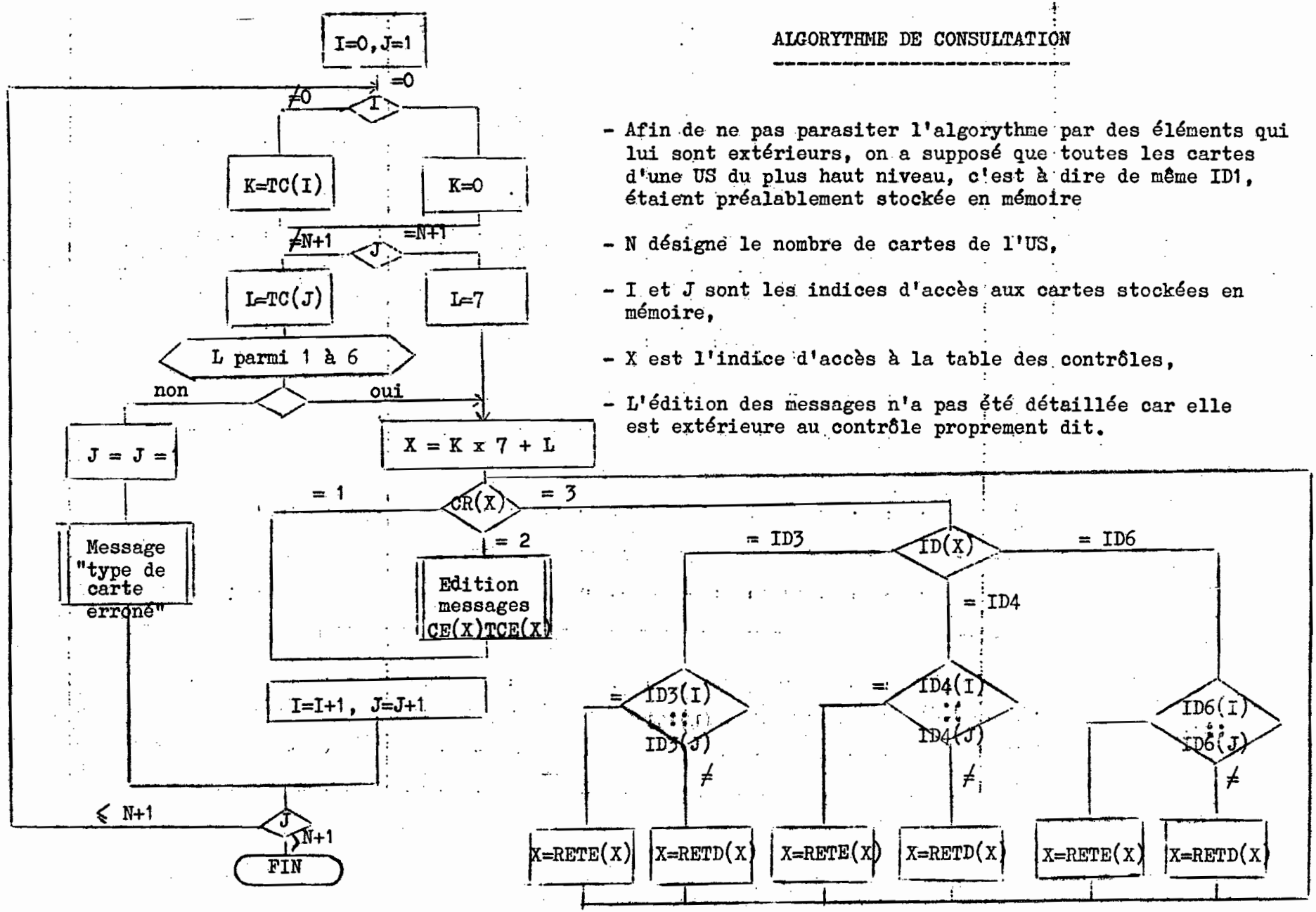
L'absence d'identifiant à un niveau mineur ne pose pas de problème ; ainsi, si les emplois ne sont pas numérotés (pas d'ID6) on notera, à la ligne 60, CR = 1 et on supprimera les lignes 62 et 63.

Remarque :

Il faut ici anticiper quelque peu sur la suite pour dire qu'en cas de détection de doubles, et si on prévoit une correction manuelle, il est souhaitable d'éliminer un des éléments du doublet, sinon les deux, afin de permettre à la mise à jour de jouer correctement son rôle. En effet la conséquence obligée de la détection de double est l'élimination de l'élément erroné. Si les deux sont conservés dans le fichier, la mise à jour ne saura lequel choisir étant donné qu'ils ont les même éléments d'identification et ne pourra que les éliminer tous les deux. En éliminer un au hasard dès la détection fera qu'on aura une chance sur deux d'avoir conservé le bon. Les éliminer tous les deux obligera à recréer le bon mais supprimera toute ambiguité quant à la validité de celui qui reste.

.../...

ALGORYTHME DE CONSULTATION



- Afin de ne pas parasiter l'algorythme par des éléments qui lui sont extérieurs, on a supposé que toutes les cartes d'une US du plus haut niveau, c'est à dire de même ID1, étaient préalablement stockée en mémoire
- N désigne le nombre de cartes de l'US,
- I et J sont les indices d'accès aux cartes stockées en mémoire,
- X est l'indice d'accès à la table des contrôles,
- L'édition des messages n'a pas été détaillée car elle est extérieure au contrôle proprement dit.

Exemple :

La séquence de cartes ci-après entraînera l'émission des messages indiqués à droite.

ID1	ID3	ID4	ID6	TC	Messages
1	-	-	-	1	
1	-	-	-	2	
1	1	-	-	3	
1	2	-	-	3	
1	2	-	-	3	CARTE 3 EN DOUBLE
1	3	-	-	A	CARTE DE TYPE A ERRONE
1	4	-	-	1	CARTE 1 DECLASSEE
1	-	1	1	6	MANQUE CARTE 4, MANQUE CARTE 5
1	-	1	1	6	CARTE 6 EN DOUBLE
1	-	2	-	4	
1	-	2	-	5	
1	-	3	-	4	
1	-	4	-	5	MANQUE CARTE 5, MANQUE CARTE 4
1		4	1	6	
1		4	2	6	
2	-	-	-	A	CARTE DE TYPE A ERRONE
2	-	-	-	B	" " " B ERRONE
2	1	-	-	C	" " " C "
2	-	1	-	D	" " " D "
2	-	1	-	E	" " " E "
2	-	1	1	F	" " " F "
					TOUS TYPES DE CARTES ERRONES

.../...

VII-2-3- Les contrôles de validité et cohérence des données

Ils ont été présenté dans le chapitre 2 sur les contrôles et corrections. D'un point de vue informatique il y a peu de choses à ajouter à ce qui y est dit. On entre là dans un domaine qui est parfaitement spécifique de chaque enquête et il n'existe pas à notre connaissance de démarche standard qui permette de définir et ordonner ce type de contrôle. Tout au plus peut on dire, ou plutôt répéter que le contrôle des variables clés devra précéder celui des variables secondaires d'une part, que les contrôles de validité ou formels, qui vérifient seulement que les données prennent bien des valeurs qui appartiennent au champ des possibles sans souci du contexte dans lequel elles se trouvent, devront précéder les contrôles de cohérence d'autre part. Ces deux objectifs peuvent entrer en conflit dans certains cas. Cela devra conduire à un réexamen de la pertinence des contrôles. Ils convient également d'insister sur l'intérêt, une fois tous contrôles faits, d'un examen global des erreurs détectées. Cette examen ne peut bien sûr être exhaustif car le nombre de configurations d'erreurs possibles est très élevé (2^n s'il y a n contrôles). Il s'agira donc de rechercher les configurations les plus significatives du point de vue de la détection des contrôles qui sont positifs à tort.

On pourra éventuellement autoriser, lorsqu'elles sont à redressement manuel, la confirmation des erreurs ou anomalies détectées. Certaines structures d'unités statistiques, qui paraissent abberantes du point de vue des ratios retenus peuvent néanmoins être réelles. Il faudra donc se donner la possibilité de dire au programme que telle erreur qu'il détecte n'en est pas une pour telle unité statistique particulière en associant au contrôle correspondant un code de confirmation normalement inactif et que le gestionnaire activera par un bordereau de correction. Il faudra manier cette possibilité avec prudence par crainte de voir des gestionnaires fatigués confirmer un peu tout et n'importe quoi.

VII-2-4- Analyse et amélioration des contrôles et redressements

VII-2-4-1- Analyse des évènements

Les indicateurs sur l'état du fichier présentés au paragraphe V ci-dessus apportent un certain nombre d'éléments d'appréciation de la pertinence des contrôles réalisés. On peut vouloir raffiner et examiner au niveau individuel les effets des redressements, notamment automatiques, en comparant l'état d'une unité statistique avant et après redressement. Une telle étude permettra de se faire une idée très précise de la qualité finale du fichier. Elle sera surtout intéressante dans le cas des enquêtes permanentes ou répétitives dont on pourra grâce à cela améliorer progressivement la qualité.

.../...

La méthode consistera alors chaque fois qu'une erreur entraînant redressement est rencontrée à verser l'état de l'US à cet instant dans un fichier "EVENEMENTS". Si une US est affectée par plusieurs redressements successifs le fichier EVENEMENTS recevra autant d'états successifs de l'US dont l'examen permettra de juger les effets des redressements, les abérrations éventuelles qu'ils provoquent dans sa structure.

Cette méthode est lourde et coûteuse. Le fichier "EVENEMENTS" risque de prendre assez rapidement une ampleur considérable. Il est de plus assez probable qu'une analyse seulement statistique des résultats sera insuffisante et qu'on sera amené à examiner, au moins en partie, les cas individuels. Ces raisons font qu'on aura intérêt, pour les enquêtes de quelque importance, à limiter l'opération à un sous-échantillon.

VII-2-4-2- Contrôle des programmes

Un moyen très efficace de contrôle d'un programme utilisant le redressement automatique consistera à lui soumettre le fichier qu'il vient de contrôler et apurer. Il devra alors ne détecter aucune erreur puisqu'il vient de corriger toutes les erreurs qu'il est chargé de détecter. S'il en va autrement cela signifie que les redressements se font mal et qu'il va donc falloir corriger le programme. Cette procédure est de peu d'intérêt pour les programmes utilisant le redressement manuel puisque ceux-ci ne modifient pas les données erronées et que le fichier de sortie est alors une image fidèle du fichier d'entrée ; que les contrôles soient pertinents ou non on obtiendra en principe le même résultat ; un résultat différent indiquera alors très précisement une erreur de programmation qui entraîne un recouvrement des données d'entrée par quelque chose d'autre.

La seule condition à cette possibilité est que le fichier de sortie ait le même dessin que le fichier d'entrée ou un dessin qui reste compatible en entrée du programme. On cherchera donc à définir dès l'entrée dans la chaîne d'apurement un dessin de fichier qui reste ensuite aussi invariant que possible au fil des différents programmes de contrôle et redressement. Ce pourra être réalisé dès la saisie lorsque celle-ci se fait sur support magnétique. Lors d'une saisie sur carte d'un fichier de structure relativement complexe cela posera bien sûr quelques problèmes et entraînera un coût supplémentaire d'analyse et programmation qui peut n'être pas négligeable mais qui sera très probablement largement amorti par la facilité qu'apportera un dessin invariant à l'écriture des différents programmes de la chaîne et à l'exploitation (notamment au fin de contrôle) des différents fichiers intermédiaires.

.../...

VIII - LA MISE A JOUR

Nous regrouperons sous le vocable "mise à jour" toutes les opérations de rapprochement de fichiers qu'on peut trouver dans la chaîne d'apurement qu'il s'agisse :

- d'une mise à jour stricto-sensu c'est à dire d'une opération sur le fichier des données de l'enquête qui consistera à ajouter, supprimer ou modifier une unité statistique,

- d'une consultation d'un fichier de données externes destinées à étayer les contrôles,

- d'une consultation ou alimentation d'un fichier directeur (cf paragraphe V ci-dessus).

Ces différentes opérations bien qu'elles aient des objectifs tout à fait distincts relèvent en effet de la même technique informatique.

Il y aura mise à jour dans toute enquête utilisant le redressement manuel, un fichier directeur ou un fichier de références externes. Selon qu'on utilisera une, deux ou les trois possibilités la mise à jour devra rapprocher deux, trois, ou quatre fichiers simultanément ou deux à deux ou trois à trois selon l'organisation et l'enchaînement des opérations. Les actions à entreprendre dans les différents cas de mise à jour varieront selon qu'on appariera le fichier des données d'enquête avec le fichier de mise à jour, le fichier directeur ou le fichier de références externes.

VIII-1- Les modes de mise à jour

Nous distinguerons le cas de la saisie sur carte de celui de la saisie sur support magnétique pour constater ensuite que les deux peuvent se rejoindre.

.../...

VIII-1-1- Saisie sur cartes

Les modifications portent habituellement sur un petit nombre des variables. Si les données de chaque unité statistique occupent plusieurs cartes en 1ère saisie il est probable qu'un petit nombre d'entre elles (le plus souvent une seule) seront modifiées par les corrections. Mieux, à l'intérieur des cartes modifiées une seule variable sera affectée dans la plupart des cas. En principe on devra tendre à saisir à nouveau, outre les données d'identification, les seules variables erronées afin de réduire le temps de travail et d'éviter qu'une variable correcte après la première saisie devienne fausse après le seconde par suite d'erreur à ce niveau. Au total nous poserons que l'unité de mise à jour est la carte perforée et nous distinguerons 3 cas.

1er cas - Mise à jour par cartes entières

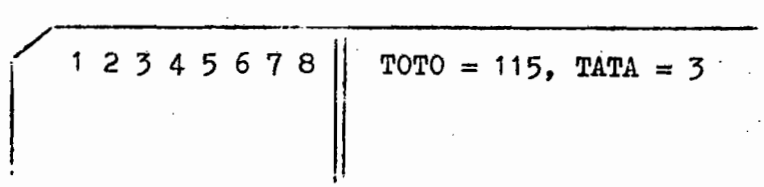
On saisira à nouveau dans son entier toute carte contenant une information erronée. Nous venons de voir l'inconvénient du procédé. Il a en contrepartie l'avantage de rendre la programmation de la mise à jour la plus simple possible puisque la nouvelle carte viendra remplacer celle qui contenait l'information erronée.

2ème cas - Mise à jour par "Zones à blanc"

Si une carte contient en colonnes 1 à 10 les identifiants de l'unité statistique et en colonnes 21 et 22 une information erronée, on perforera une carte de mise à jour qui contiendra les identifiants et en colonnes 21 et 22 la valeur de correction. Le reste de la carte sera laissé à blanc. L'algorithme de mise à jour comparera position à position les cartes origine et correction et corrigera celles de la première par les valeurs des positions correspondantes de la seconde qui ne sont pas à blanc. La programmation bien qu'un plus complexe que dans le cas précédent reste malgré tout assez simple. Par contre le risque d'erreur dans le positionnement des variables à corriger est, on l'a constaté, assez important.

3ème cas - Mise à jour par identification des variables

On identifie cette fois les variables à corriger par un nom de code qu'on fera explicitement figurer dans la carte correction qui pourra alors se présenter comme ceci :



.../...

Dans l'unité statistique de numéro 12345678 on corrige les variables TOTO et TATA par les valeurs 115 et 3 respectivement.

Cette 3ème méthode, plus fiable que les précédentes, à l'inconvénient de compliquer considérablement la programmation.

Après les avoir employées toutes trois et constatées que les erreurs engendrées par la saisie à nouveau étaient très peu nombreuse nous recommanderons l'usage de la première qui à l'avantage de la simplicité.

VIII-1-2- Saisie sur support magnétique

C'est la naturellement la mise à jour par identification des variables qui s'impose comme la seule satisfaisante. On a en effet dans ce cas défini et programmé des formats de saisie qui identifient toutes les variables du fichier afin de guider la saisie. Ne pas s'en servir pour les mises à jour serait absurde. De plus les enregistrements de saisie sont à priori plus longs puisqu'on est plus contraint par la dimension du support ; l'inconvénient, relativement mineur dans le cas de la carte perforée, qu'il y a saisir à nouveau des informations déjà corrects devient de ce fait redhibitoire. Le programme de saisie aura alors pour mission de replacer les variables modifiées à leur juste place dans l'enregistrement de telle manière qu'au niveau du programme de mise à jour on soit ramené au cas de mise à jour par "zones à blanc".

VIII-2- Codes de mise à jour et mise à jour par le contexte

Les cas de mise à jour sont la création la suppression et la modification.

La modification s'applique aux variables.

La création et la suppression peuvent s'appliquer aux unités statistiques, quelque soit leur niveau, ou aux cartes si la saisie est faite sur ce type de support.

La mise à jour peut être clairement précisée par un code ad hoc qui précisera le type d'opération qu'on veut réaliser (par exemple : C = création, S = suppression, M = modification).

.../...

Elle peut également être définie par le contexte en posant les règles suivantes :

1 - si un enregistrement de mise à jour n'a pas de correspondant de même identifiant dans le fichier permanent cela signifie qu'il s'agit d'une création,

2 - s'il a un correspondant dans le fichier permanent on devra distinguer deux sous cas :

21 - tout l'enregistrement, en dehors des zones d'identification, est à blanc ; cela signifie qu'il s'agit d'une suppression,

22 - l'enregistrement n'est pas entièrement à blanc ; cela signifie qu'il s'agit d'une modification et on corrigera les variables renseignées dans l'enregistrement de mise à jour,

3 - si le fichier est hiérarchisé toute suppression de l'enregistrement d'une unité statistique majeur entraînera la suppression de toute sa descendance.

Ces règles suffisent à définir la mise à jour par le contexte d'un fichier saisie sur support magnétique car il prend dès la saisie sa forme définitive en terme de structure. En ce qui concerne les fichiers saisis sur carte le fait qu'une unité statistique puisse être répartie sur plusieurs cartes les parasitent quelque peu du moins pour ce qui concerne la suppression. Voyons à partir de l'exemple du paragraphe VIII-2-2- sur les contrôles de structure les cas de suppression qu'on peut envisager.

Rappelons que :

- le "ménage-logement" occupe les cartes de types 1 et 2 et qu'il est identifié par ID1,

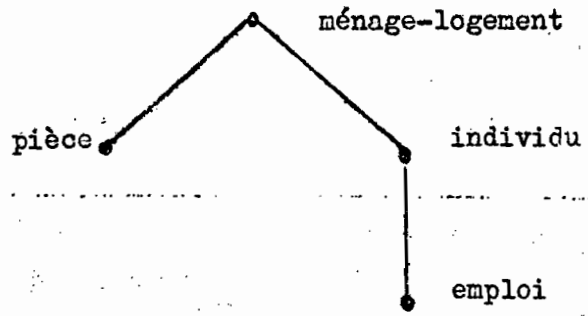
- la "pièce du logement" occupe une carte de type 3 et qu'elle est identifiée par ID3,

- l'individu occupe les cartes de types 4 et 5 et qu'il est identifié par ID4,

- l'emploi occupe une carte de type 6 et qu'il est identifié par ID6.

.../...

Le structure du fichier est représentée par l'arbre suivant :



Le tableau ci-dessous recense les cas de suppression possibles dans le cas où le contrôle de structure élimine automatiquement les cartes en doubles (cf VII-2-2-2-3- remarque de fin de paragraphe). La mention R signifie que l'identifiant est renseigné, son absence qu'il ne l'est pas. Les zones données des cartes sont dans tous les cas à blanc. Toute autre configuration que celles énumérées est en principe invalide.

.../...

CAS	ID1	ID3	ID4	ID6	Type de carte	Cas de suppression
1	R	-	-	-	-	Suppression du ménage identifié par ID1 et de toute sa descendance
2	R	R	-	-	3	Suppression de la pièce identifié par ID3
3	R	-	-	-	3	Suppression de toutes les pièces du ménage-logement identifié par ID1
4	R	-	R	-	4	Suppression de l'individu identifié par ID4 et de sa descendance
5	R	-	-	-	4	Suppression de tous les individus du ménage identifié par ID1
6	R	-	R	R	6	Suppression de l'emploi identifié par ID6
7	R	-	R	-	6	Suppression de tous les emplois de l'individu identifié par ID4

Remarque :

La présence du type de carte dans les cas 2, 4 et 6 est redondante. Elle peut permettre un contrôle complémentaire.

.../...

Choisir entre mise à jour par codes et mise à jour par le contexte est difficile. La première méthode offre en principe une sécurité complémentaire qui peut être contrebalancée par des erreurs dans la manipulation des codes. Nous ne trancherons pas.

VIII-3- Les erreurs de mise à jour

Les erreurs de mise à jour proprement dite sur le fichier des données sont au nombre de trois ; ce sont :

- la création d'un existant,
- la suppression d'un non-existant,
- la modification d'un non-existant.

La tentative de création d'un existant ne sera détectée que dans la mise à jour par codes ; dans la mise à jour par le contexte elle sera considérée comme une modification portant sur la totalité des données de l'enregistrement incriminé.

Ces 3 cas pourront se détailler plus ou moins selon le niveau auquel ils interviennent. Dans le cas de tentative de suppression d'un non-existant par exemple on pourra préciser le niveau de l'US concerné et, s'il y a lieu, le type de carte.

On devra prendre en compte également, en cas de mise à jour par le contexte les configurations de suppression non valides (voir tableau du paragraphe VIII-2- ci-dessus).

En tout état de cause le nombre de cas d'erreurs de mise à jour est bien déterminé et la tentation peut être grande de définir des textes de messages standards et généraux qui s'appliquent à toutes les enquêtes. Nous le déconseillons formellement. Les erreurs de mise à jour sont en effet parmi les plus difficiles à appréhender dans toutes leurs subtilités par les gestionnaires d'enquêtes peu au fait des concepts informatiques utilisés. Là, plus qu'ailleurs encore, il faudra se montrer clair, précis et concret.

La présence d'un fichier directeur permettra d'affiner quelque peu le contrôle de la mise à jour en signalant par exemple qu'on tente de corriger ou supprimer une unité statistique qu'on avait décrétée correcte lors d'un passage précédent, etc...

.../...

VIII-4- Les sorties de la mise à jour

La mise à jour prend en compte le fichier permanent d'enquête tel qu'il était après le précédent passage de la chaîne d'apurement et un fichier mouvement qui contient les US nouvellement saisie et les corrections de tout ou partie de celles pour lesquelles on a détecté des erreurs lors des précédents passages. On aura le plus souvent une partie seulement des corrections car si les passages de la chaîne d'apurement se font à intervalles de temps fixe, ce qui est souhaitable, il est fort peu probable que les gestionnaires auront corrigés au moment d'un nouveau passage toutes les erreurs détectées lors du précédent.

Les contrôles font suite à la mise à jour. Si on leur soumet à chaque tour la totalité du fichier permanent on va pour les unités statistiques qui n'ont pas été corrigées sortir plusieurs fois de suite les mêmes messages d'anomalies ce qui risque de produire des conflits avec l'atelier de gestion notamment lorsqu'il s'agira d'erreurs qui ont bien été traitées par les gestionnaires mais trop tard pour être prises en compte dans le passage compte-tenu des délais de saisie et de transmission. C'est pourquoi il parait de beaucoup préférable de ne soumettre à contrôle que les unités statistiques affectés par la mise à jour. On évitera ainsi toute redondance. On complètera ce dispositif par des passages à longs intervalles sur l'ensemble du fichier destinés à draîner les erreurs résiduelles oubliées par les gestionnaires. On placera ces passages - bilan à la fin de l'apurement ou on en décidera au vue des statistiques sur l'état d'avancement de l'apurement fournies par le fichier directeur.

VIII-5- Reprise des mises à jour

Ce qui suit ne vaut que pour la mise à jour proprement dite et n'a donc d'intérêt que lorsqu'on utilise les redressements manuels. Les états successifs d'une unité statistique peuvent alors être relativement nombreux du fait des itérations de la chaîne d'apurement. Lors d'un premier passage l'unité statistique est entièrement saisie puis contrôlée. Le contrôle peut provoquer l'émission de messages d'erreurs qui entraîneront de la part du gestionnaire la rédaction d'un bordereau de correction qui viendra modifier le premier état de l'unité statistique. Celle-ci, dans son nouvel état, sera à nouveau contrôlée. Si la correction a été mal faite il y aura à nouveau détection d'erreur, rédaction d'un bordereau de correction, modification de l'état de l'unité statistique, contrôle de ce nouvel état, etc... On a vu dans certaines enquêtes, pour des unités statistiques présentant des cas d'erreurs particulièrement difficiles, le processus se répéter jusqu'à quatre ou cinq fois (il convient toutefois de préciser que les règles de définition et ordonnancement des contrôles présentées dans les paragraphes précédents n'avaient pas été strictement respectées !).

.../...

Au total l'état du fichier dans la phase d'apurement est, à un instant donné, la résultante d'un certain nombre de passages dont chacun a introduit un certain nombre d'unités statistiques, on a supprimé un certain nombre d'autres qu'on y avait chargé à tort, on a modifié d'autres pour lesquelles on avait détecté des erreurs. A cet instant toute US figurant au fichier a été contrôlée au moins une fois lors de son introduction et éventuellement modifié une ou plusieurs fois par des corrections. Cesévènements ont entraîné des modifications corrélatives du fichier directeur qui est chargé de les enregistrer.

Mais si le fichier directeur et le fichier des données renseignent complètement sur l'état des travaux après le dernier passage ils ne conservent pas une trace complète des évènements qui, au fil des passages, ont affectés les unités statistiques. Cette absence d'historique peut se révéler génante. Notamment dans le cas où on détecte après plusieurs passages de la chaîne d'apurement, une erreur programmation entraînant la perte de certains enregistrements ou dans le cas de destruction accidentelle du fichier. Il sera alors très difficile, voire impossible de remonter le fil des évènements de manière à détecter les unités statistiques ainsi supprimée puis de retrouver les données qui les concernent dans les lots de saisies. On risque fort d'être amené à refaire tous les passages de contrôle déjà réalisés pour retrouver un fichier correct ou identique au fichier perdu.

Pour pallier ces difficultés on pourra stocker les lots successifs de 1ère saisie et mise à jour dans un fichier historique en les datant par un numéro d'ordre séquentiel croissant : le 1er lot portera le n° 1, le second le n° 2, etc... Si une unité statistique a été saisie pour la 1ère fois au 1er passage, puis modifiée aux 2ème, 3ème et 4ème, il suffira de trier le fichier archives sur les identifiants de l'US en majeur et le n° de passage en mineur pour avoir à la suite les uns des autres ses différents états. Si l'algorithme de mise à jour est conçu de manière à admettre plusieurs modifications successives d'un enregistrement lors d'un même passage, on pourra en un seul passage utilisant le fichier archive comme fichier de mise à jour reconstituer l'état du fichier des données avant sa perte ou corrigé des erreurs de gestion qu'on vient de détecter. Le contrôle proprement dit ne sera donc à refaire que sur le fichier ainsi reconstitué.

.../...

VIII-6- Organisation des programmes de mise à jour

La mise à jour est un filon inépuisable pour les épreuves des examens de programmation. En effet les problèmes que pose la cinématique des fichiers dès lors que leur nombre est supérieur à deux où qu'ils prennent des structures hiérarchiques complexes deviennent rapidement difficiles à résoudre. Il en résulte, en l'absence d'une méthode standardisée de gestion des lectures, des algorithmes de mise à jour très divers d'un auteur à l'autre, souvent fantaisiste et très difficile à comprendre et à manier pour la lecteur quand ce n'est pas pour l'auteur lui-même.

Le principe de la mise à jour séquentielle (ou appariement de fichiers) consiste à mettre en concordance des enregistrements ou groupes d'enregistrements relatifs à une même unité statistique stockés sur des supports distincts. Les enregistrements sont, sur les différents supports, rangés dans le même ordre et le fait qu'il concernent la même unité statistique reconnu par le fait que certains codes, qu'on nomme identifiants, prennent la même valeur sur les différents supports. L'identifiant est une variable discriminante qui prend une valeur différente pour chacune des unités statistiques du champ étudié. La mise à jour est aisée lorsque le nombre de fichiers est limité à deux, que chaque unité statistique est représentée par un et un seul enregistrement dans chaque fichier et que ceux-ci ne sont pas hiérarchisés. Les difficultés apparaissent et croissent très rapidement lorsque :

- le nombre de fichiers à apparier devient supérieur à deux,

- les fichiers se hiérarchisent c'est à dire lorsque apparaissent des unités statistiques de différents niveaux emboitées les unes dans les autres,

- certaines unités statistiques ne sont pas représentées dans tous les fichiers,

- à l'inverse lorsque certaines unités statistiques sont représentées par plusieurs enregistrements décrivant ses états successifs dans un fichier mouvement,

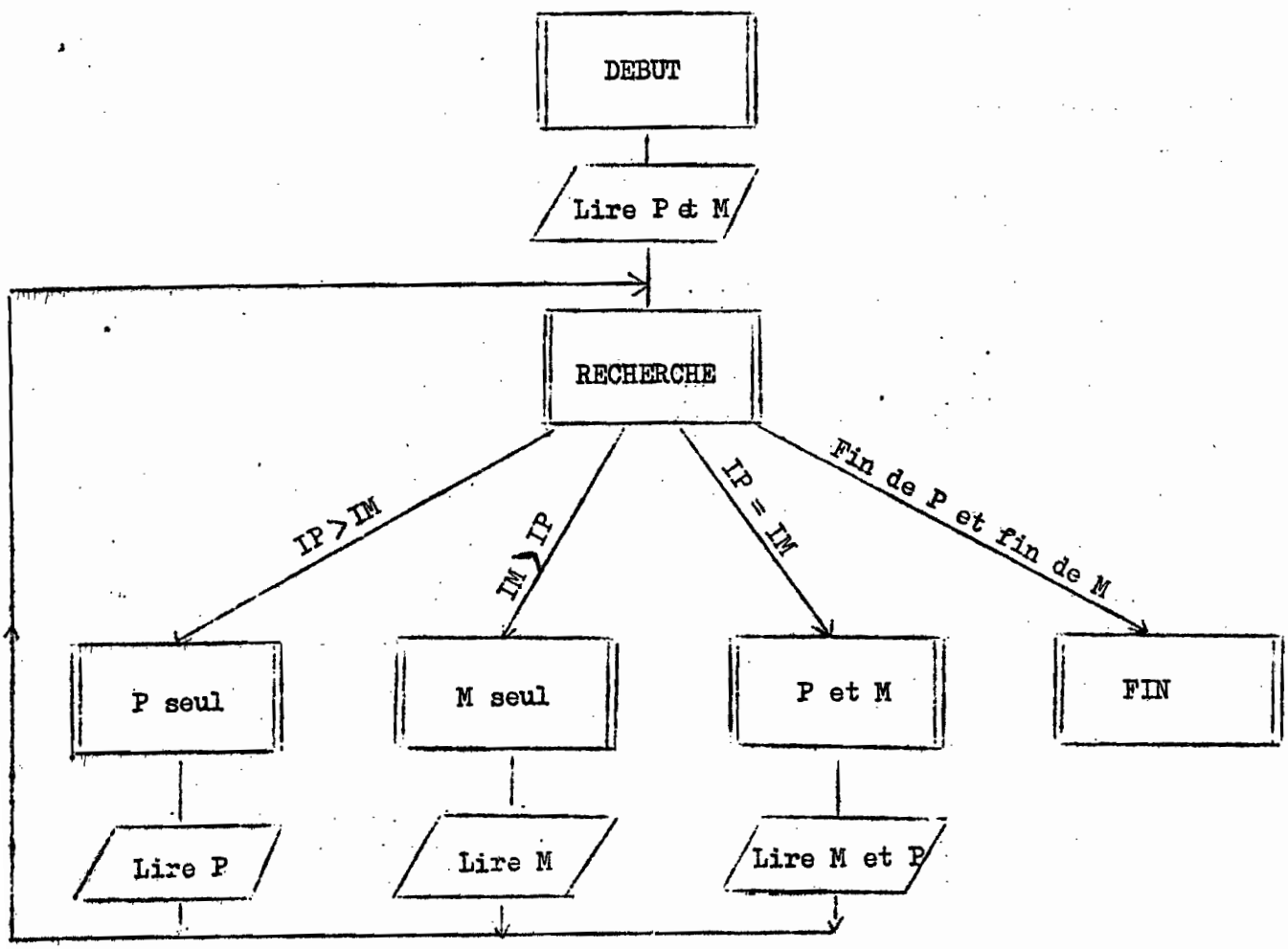
- lorsqu'enfin on cherche à apparier des fichiers contenant des unités statistiques qui ne sont pas de même niveau (un fichier de communes avec un fichier d'individus par exemple).

Il n'en reste pas moins que les problèmes de la mise à jour sont bien connus et définis et qu'on doit pouvoir leur trouver des solutions également définies. C'est ce que nous allons tenter de faire en présentant un algorithme de mise à jour que nous appliqueront d'abord en cas de deux fichiers non hiérarchisés pour voir ensuite comment il peut se transposer à des cas plus complexes.

.../...

VIII-6-1- Cas de deux fichiers non hiérarchisés

La structure générale du programme sera la suivante :



On désigne par : - P le fichier permanent,
- M le fichier mouvement,
- IP l'identifiant sur P,
- IM l'identifiant sur M.

.../...

La dynamique du programme est gérée par un module, nommé RECHERCHE, auquel on soumet chacun des couples d'enregistrements qui se forment au fil des lectures des deux fichiers et qui a pour mission d'orienter vers le module de traitement de chacun des 4 cas spécifiques qu'on peut rencontrer ici à savoir :

- P seul : l'enregistrement du fichier permanent n'a pas de correspondant dans le fichier mouvement,

- M seul : l'enregistrement du fichier mouvement n'a pas de correspondant dans le fichier permanent,

- P et M : l'enregistrement du fichier mouvement et l'enregistrement du fichier permanent correspondent à la même unité statistique,

- FIN : le fichier permanent et le fichier mouvement sont tous deux terminés, il faut donc exécuter les opérations de fin de programme. Il faut noter que si l'on rencontre la fin du fichier mouvement avant celle du fichier permanent on se trouve dans le cas Pseul ; à l'inverse on se trouve dans le cas Mseul.

Nous reviendrons plus en détail sur la structure du module de RECHERCHE.

Quelques soient les traitements réalisés par les modules "Pseul", "Mseul" et "P et M" ils se terminent obligatoirement par une lecture du ou des fichiers dont ils viennent de traiter l'enregistrement afin de fournir à RECHERCHE le nouveau couple qu'il va falloir analyser. Les fonctions propres de ces 3 modules varieront selon le cas de figure dans lequel on se trouve. Ainsi :

- Pseul aura, dans le cas d'une mise à jour stricto-sensu, pour seul objectif de recopier l'enregistrement lu sur le fichier permanent de sortie car il correspond alors à une unité statistique qui n'est pas affectée par la mise à jour,

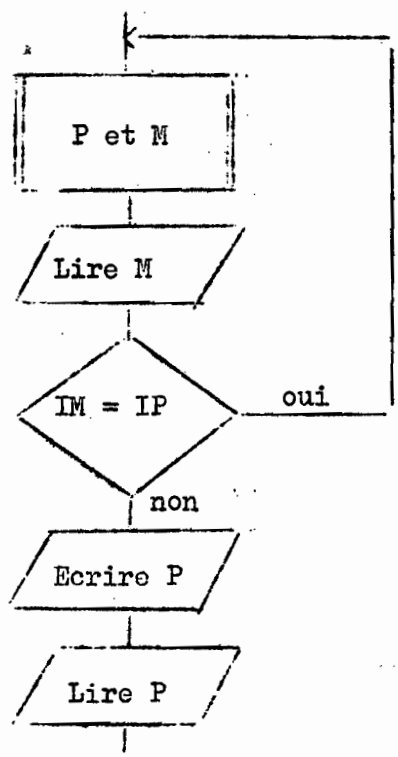
- Mseul correspond, dans le cas de mise à jour par le contexte, à une création, l'enregistrement sera donc versé au fichier permanent et soumis à contrôler dans le corps même du module ou ultérieurement dans les programmes suivants de la chaîne d'apurement. Si on utilise un code de mise à jour on cherchera s'il n'y a pas conflit entre celui-ci et le cas de figure (suppression ou modification d'un non existant au fichier permanent),

- P et M correspond aux cas de modification ou suppression que le module devra préciser à partir du code de mise à jour ou de la configuration de l'enregistrement du fichier mouvement.

.../...

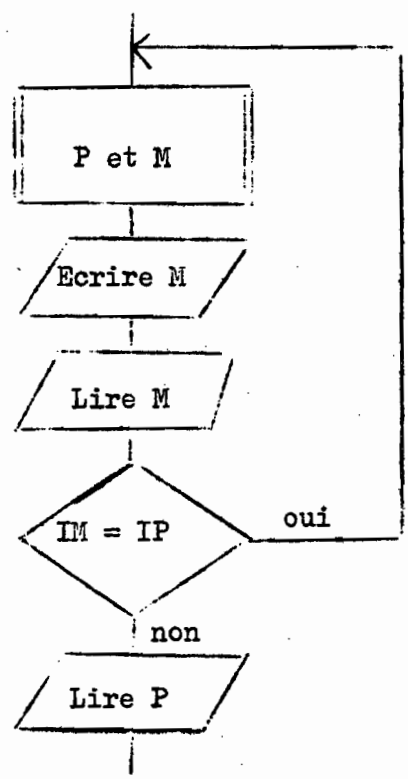
L'organigramme ci-dessus correspond au cas ou à un enregistrement du fichier permanent ne peut correspondre qu'un enregistrement du fichier mouvement et réciproquement. Si l'un des deux, le fichier mouvement par exemple, peut contenir plusieurs enregistrements en correspondance avec un seul du fichier permanent la structure du module P et M s'en trouvera modifiée, mais pas celle de Pseul et Mseul. On distinguera deux cas :

1 - le fichier mouvement est de type historique (voir VIII-5 ci-dessus). Il contient donc plusieurs mises à jour successives de l'enregistrement du fichier permanent et on veut reconstitués l'état de celui-ci après prise en compte de toutes les mises à jour dans l'ordr dans lequel elles ont été demandées. La structure de P et M sera alors la suivante :



.../...

2 - le fichier P, contient des unités statistiques de rang supérieur à celles de M. C'est par exemple un fichier de références communales qui va servir à contrôler ou alimenter les enregistrements individus de M. Il faudra alors différer la relecture de P jusqu'à ce qu'on ai défilé tous les enregistrements de M en correspondance avec son enregistrement en cours. La structure du module devient :

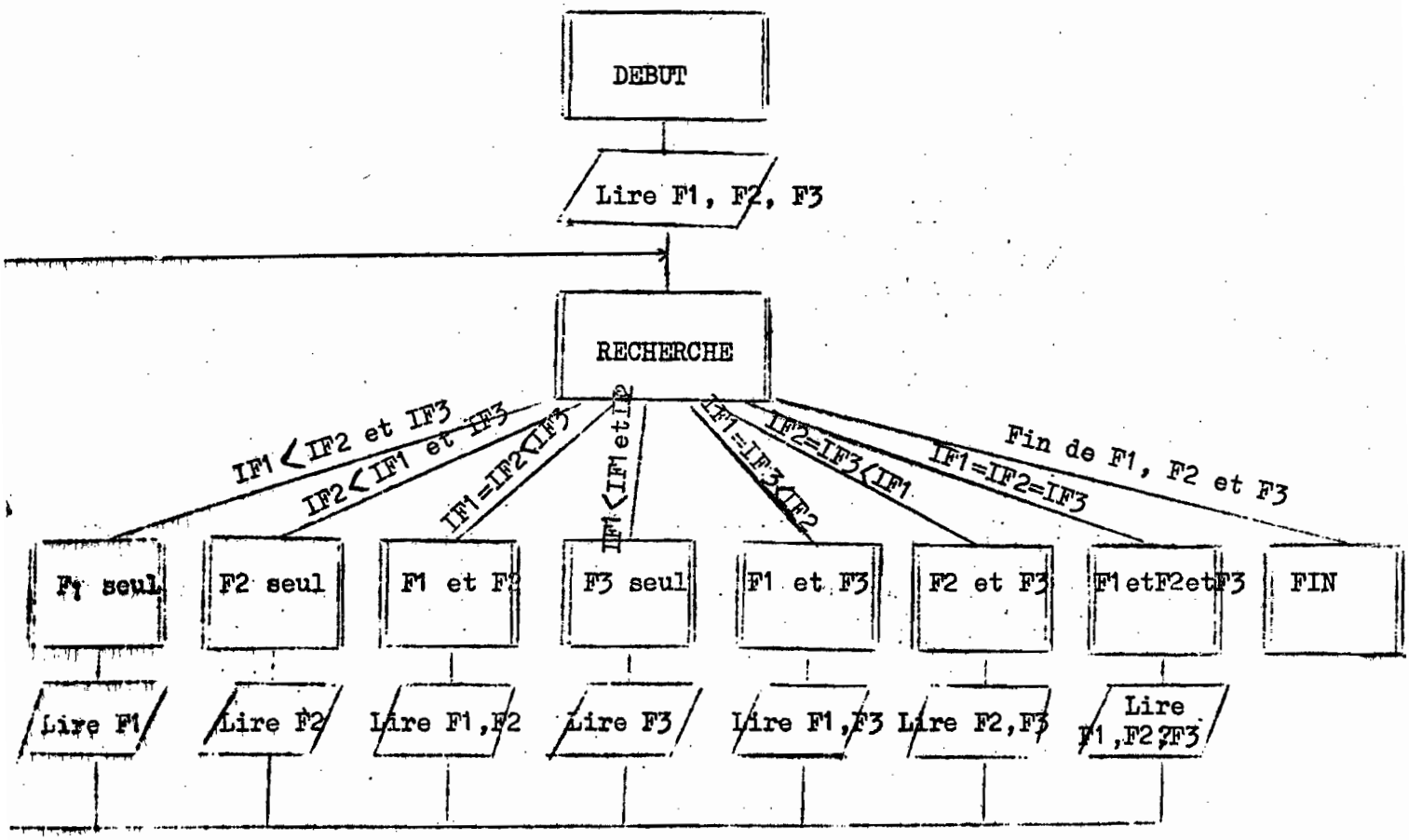


.../...

VIII-6-2- Extension à plusieurs fichiers

La structure générale du programme reste la même, dans le cas de fichiers non hiérarchisés, quelque soit le nombre de fichiers pris en compte. Seul augmente le nombre de cas d'appariements à envisager qui est égal à 2^n (2^{n-1} cas d'appariements effectifs plus le cas de fin de tous les fichiers) si n est le nombre de fichiers.

Pour trois fichiers F1, F2 et F3 d'identifiants IF1, IF2 et IF3 on aura la structure suivante :



Les remarques du paragraphe précédent relatives aux fins de fichiers non simultanées et au fait qu'un fichier peut contenir des multiples par rapport à l'autre s'appliquent ici. Nous ne développerons pas les cas possibles qui ne déduisent aisément de ce qui précède.

.../...

VIII-6-3- Cas de fichiers hiérarchisés

Revenons au cas de deux fichiers, mais en supposant cette fois qu'ils sont hiérarchisés. Il apparaît clairement que la structure de programme telle que nous venons de la définir n'est plus suffisante du fait qu'on a maintenant à faire à plusieurs types d'unités statistiques identifiées par autant d'identifiants et auxquelles correspondent autant de types d'enregistrements. Les choses se compliquent encore dans le cas d'une saisie sur cartes perforées car alors les données d'une unité statistique peuvent être réparties sur plusieurs enregistrements-cartes différents. Comment notre algorithme peut-il se modifier pour prendre en compte ces faits nouveaux on peut envisager deux solutions :

- la 1ère s'en tient à la structure des fichiers en considérant successivement les différents niveaux d'unités statistiques rencontrés. En s'en tenant d'abord aux unités statistiques du plus haut niveau on retrouve dans un 1er temps un algorithme identique à celui qui traite les fichiers non hiérarchisés. En faisant agir le module RECHERCHE sur l'identifiant majeur on aboutira à l'un des 3 cas connus Pseul, Mseul et P et M. Pour ce qui est de Pseul et Mseul les traitements diffèreront peu de ce que nous avons déjà dit car il s'agira alors dans le cas banal de recopier plusieurs enregistrements au lieu d'un seul. Pour ce qui est de P et M par contre il va falloir déterminer à quel niveau s'opère les opérations de suppressions et (ou) modifications qui sont demandées et affiner l'analyse. On pourra considérer que l'ensemble des enregistrements de l'unité statistique majeur en cours de traitements constituent dans le fichier permanent et le fichier mouvement des sous-fichiers logiques d'unités statistiques du 2ème niveau auxquels on pourra appliquer en principe un algorithme identique à celui du 1er niveau mais travaillant sur l'identifiant du 2ème niveau. S'il existe un 3ème niveau d'US on pourra tenir le même raisonnement entre 2ème et 3ème niveau et ainsi de suite. L'algorithme semble donc devoir se développer selon une loi bien définie mais il deviendra rapidement complexe et difficile à manier. De plus se poseront, du fait de la saisie sur cartes ou d'une hiérarchie de fichier à plusieurs branches par exemple, un certain nombre de problèmes que nous n'avons pas recensé ici et qui viendront encore obscurcir l'économie générale du système. C'est pourquoi il nous parait préférable de renoncer à cette solution dont la mise au point est difficile.

- la 2ème solution permet de se ramener au cas des fichiers non hiérarchisés à partir d'une manipulation des identifiants. Reprenons une fois encore l'exemple du fichier ménages-logements saisi sur carte que nous avons déjà utilisé à deux reprises. On y constate que dans tout enregistrement-carte relatif à une unité statistique mineure sont reportés les identifiants de ses ascendants. On a par ailleurs recommandé afin de faciliter le tri du fichier avant le contrôle de structure de "sortir" ces identifiants de la carte proprement dite pour les placer en tête de manière à ce que les positions de stockage des identifiants mineurs existent et soient laissées à blanc dans les enregistrements des US majeurs pour lesquels ils sont sans objet. La mise à jour précédant tous les contrôles cette opération et ce tri devront en fait être réalisés avant elle.

.../...

Toute carte est complètement identifié par un super identifiant qui est la réunion des identifiants, renseignés ou non, des différents niveaux et du type de carte. Si la mise à jour prend en compte ce super identifiant sans se soucier de ses composants on réalisera une mise à jour carte à carte qui, ignorant la structure logique du fichier, nous ramènera à un appariement de fichiers non hiérarchisés. La structure du programme reste identiquement simple quelque soit la complexité de la structure des fichiers. Le procédé est donc très performant. Par contre il interdit qu'on intègre aucun contrôle au programme de mise à jour. Le premier de ceux-ci est en effet le contrôle de structure qui exige que soit stocké en mémoire, sinon la totalité des enregistrements d'une US majeur du moins tous ses identifiants et types de cartes ou d'enregistrements. Or le programme de mise à jour ainsi conçu ignore la notion d'US pour ne connaitre que celle d'enregistrement et rejette ceux-ci sur les fichiers de sortie au fur et à mesure qu'ils sont traités. Il sera donc limité à sa seule fonction propre ce qui est sans doute une bonne chose pour la clarté d'ensemble de la chaîne d'apurement.

VIII-6-4- Le module de RECHERCHE

Le module de RECHERCHE reçoit le dernier enregistrement lu de chacun des fichiers. Il a pour mission de détecter le cas d'appariement qui se présente et d'orienter vers le module de traitement afférent. On conçoit qu'il soit simple dans le cas de deux fichiers. Faute de méthode il peut devenir complexe quand on passe à trois fichiers ou davantage. Nous proposons un algorithme que nous allons appliquer aux cas de deux puis trois fichiers pour voir ensuite qu'il s'étend sans difficultés à n.

Cas de deux fichiers :

- on désigne les fichiers par F1 et F2 et leurs identifiants par IF1 et IF2,

- on utilise deux variables intermédiaires :

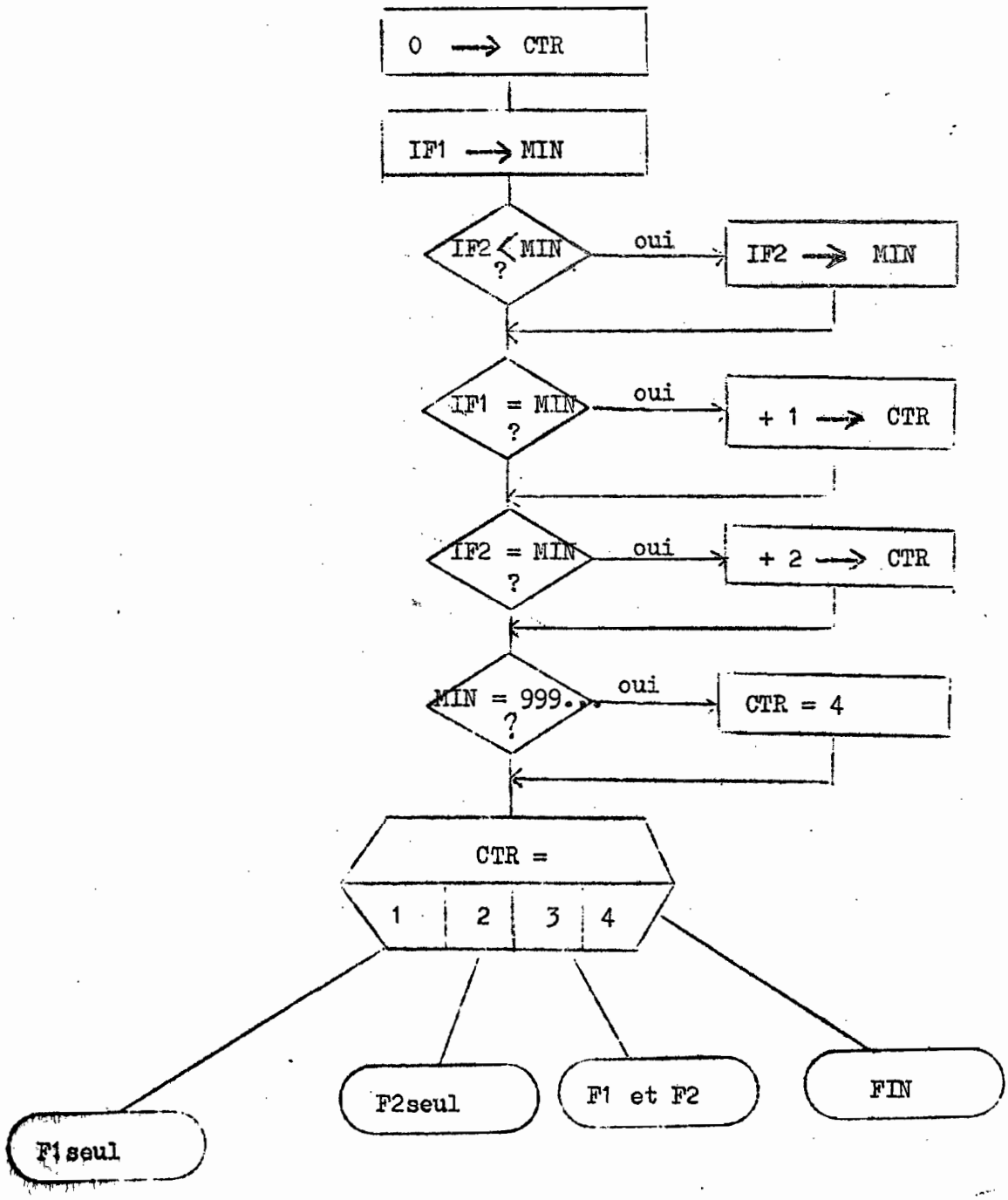
  + CTR qui est un compteur,

  + MIN qui est une zone de stockage de l'identifiant le plus petit,

- Aux ordres de lecture, qui sont hors du module, sont associées des clauses de fin qui provoqueront le chargement d'un padding dans l'identifiant du fichier lu (999 ... ou HIGH-VALUE).

Ceci étant précisé l'algorithme se présente comme suit :

.../...

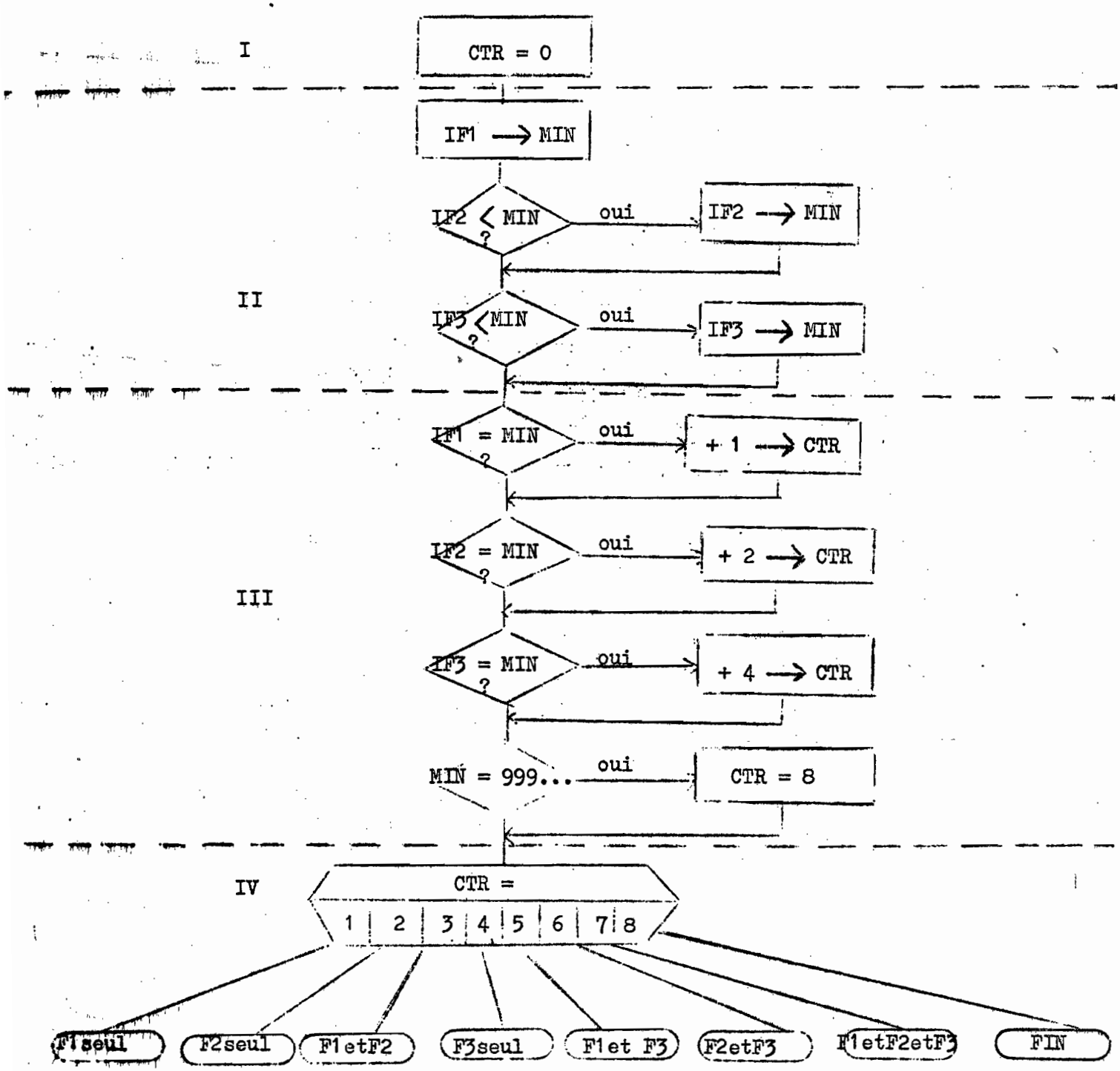


On voit que le principe consiste à charger dans MIN la valeur du plus petit des 2 identifiants (ou des deux s'ils sont égaux) puis à le comparer chacun d'entre eux en incrémentant CTR de 2^0 puis 2^1 quand le test est positif. Si MIN, c'est à dire les deux identifiants puisqu'il contient la valeur du plus petit, est égal à la valeur de padding on charge 2^2 dans CTR. La valeur de CTR permet ensuite d'orienter vers le cas d'appariement à traiter.

.../...

Cas de trois fichiers

Soit F1, F2, F3 d'identifiants IF1, IF2, IF3



.../...

On voit se détacher nettement les 4 parties de l'algorithme :

I - Initialisation du compteur,

II - Recherche du plus petit,

III - Recherche du cas d'appariement,

IV - Orientation.

Généralisation

Elle est évidente. Si n est le nombre de fichiers :

- la 1ère partie (mise à zéro du compteur) est invariable,

- la 2ème partie (recherche du plus petit) commence toujours par le chargement de IF1 dans MIN qui est suivi de n-1 tests de comparaison de MIN aux identifiants des n-1 autres fichiers. Lorsque IFi est inférieur à MIN, il est chargé dans celui-ci. Au bout du compte MIN contient la valeur du ou des plus petits identifiants des enregistrements en cours de traitement. Il contient la valeur de padding si tous les fichiers sont terminés,

- la 3ème partie (recherche du cas d'appariement) est composé de n+1 tests dont :

+ les n premiers sont une comparaison de IFi (i variant de 1 à n) à MIN ; l'égalité provoque l'incrémentation de CTR d'une valeur égale à 2^{i-1},

+ le dernier est une comparaison de MIN à la valeur de padding qui en cas d'égalité provoque le chargement de 2^n dans CTR,

- la 4ème partie (orientation) est un test aiguillage à 2^n sorties qui oriente vers les 2^n cas d'appariement possibles.

IX - LE POLISSAGE

Lorsque le fichier est complètement apuré des erreurs à redressement manuel il peut néanmoins, à propos de celles-ci subsister des anomalies minimes sans incidence au Plan Statistique mais qui peuvent néanmoins nuire au bon équilibre comptable des tableaux qu'on se propose de produire. Ce sera vrai surtout des contrôles de balance portant sur les variables quantitatives pour lesquelles on n'emettra, habituellement, un message d'anomalie que lorsque l'écart constaté entre la variable somme et le total de ses parties dépasse un certain seuil en valeur absolue ou relative ; s'il est faible l'écart subsistera dans le fichier parcequ'on juge inutile d'emettre un message dont la correction n'améliorerait pas la qualité statistique du fichier. Ce sera vrai également des anomalies qui auront été confirmées. Il convient d'effacer ces scories du fichier avant de la codifier et de le tabuler. L'opération qu'on nommera POLISSAGE, consistera donc à contrôler de manière systématique tous les équilibres comptables de chaque unité statistique et à corriger automatiquement de la manière la plus simple possible. Le raffinement ici n'est pas de mise puisque, encore une fois, ces erreurs résiduelles sont sans incidences statistiques. Afin de s'en assurer on pourra d'ailleurs sur un 1er lot contrôlé faire une tabulation de contrôle avant et après polissage et comparer les résultats.

.../...

X - SYNTHESE

Nous venons d'examiner dans le détail les différentes fonctions qui peuvent être utilisées dans une chaîne d'apurement d'enquête. Ce sont :

- le carte à bande,
- l'impression différée,
- le tri-fusion,
- la production d'histogrammes,
- la production de tableaux de contrôles,
- la production d'indicateurs, qui implique un fichier directeur, qu'il faut distinguer de la simple ...,
- production de compteurs au fil des programmes,
- la mise à jour,
- les contrôles de structures,
- les contrôles de validité et cohérence des données,
- la production de messages d'erreurs ou anomalies,
- l'amélioration des contrôles,
- le polissage.

Rares seront les enquêtes pour lesquelles on les emploiera toutes. Cela dépendra du degré de fiabilité auquel on veut aboutir qui est lié à l'importance ou à la pérennité de l'enquête. Cela dépendra aussi de ce qu'on utilisera ou non certains moyens, selon la nature des unités statistiques, le matériel disponibles, la connaissance à priori du champ enquêté. Nous distinguerons quatre possibilités qui sont :

- l'utilisation des redressements manuels,
- l'utilisation d'un fichier de références externes à l'enquête,
- la structuration du fichier d'enquête,
- la saisie directe sur support magnétique.

.../....

Par combinaison de ces possibilités on définira 16 configurations de base de la chaîne d'apurement chacune pouvant se subdiviser selon qu'on décide d'utiliser où non telle ou telle fonction complémentaire dont le seul intérêt est d'accroître le degré de fiabilité de l'enquête (ce qui n'est pas négligeable). La table de décision ci-après recense les configurations de base et les fonctions qui sont, obligatoirement ou non, associés à chacune d'elle. Elle appelle les commentaires suivants :

- le carte à bande est naturellement lié à la saisie sur cartes perforée,

- l'impression différée n'est obligatoire que lorsqu'on utilise le redressement manuel car on a alors la certitude (sauf si l'effectif enquêté est très restreint) qu'il faudra imprimer un important volume de papier. Dans les autres cas on y aura où non recours selon l'utilisation qu'on fera de certaines fonctions facultatives telles que la production d'histogrammes et de tableaux de contrôle,

- le tri-fusion, l'émission de compteurs, les contrôles de validité et cohérence des données sont toujours obligatoires,

- à l'inverse histogrammes, tableaux de contrôles, indicateurs et amélioration des contrôles sont toujours facultatifs car ce sont des fonctions qui ont pour seul objet d'améliorer la fiabilité de l'enquête,

- la mise à jour est obligatoire dès qu'on utilise le redressement manuel ou un fichier de références externes ; dans les autres cas on devra y avoir recours à un niveau dégradé si on utilise un fichier directeur,

- le contrôle de structure est naturellement lié au fait que le fichier est structuré,

- de même que l'émission de messages et le polissage sont liés au redressement manuel.

Nous allons maintenant, afin de mieux préciser la façon dont les différentes fonctions se raccordent les unes aux autres, étudier en détail les deux configurations extrêmes (1 et 16), c'est à dire la plus simple et la plus complexe, pour en déduire dans chaque cas l'organigramme fonctionnel de la chaîne d'apurement. Nous n'approfondirons pas les autres configurations qui s'en déduisent aisément.

Nota : Dans la table de décision on a noté par :

[X]	une fonction obligatoire,
[(X)]	une fonction facultative,
[ ]	une fonction absente.

.../...

LES CONFIGURATIONS DE BASE DE LA CHAINE D'APUREMENT

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Redressement manuel	N	N	N	N	N	N	N	N	O	O	O	O	O	O	O	O
Fichier de référence	N	N	N	N	O	O	O	O	N	N	N	N	O	O	O	O
Fichier structuré	N	N	O	O	N	N	O	O	N	N	O	O	N	N	O	O
Saisie sur bande	N	O	N	O	N	O	N	O	N	O	N	O	N	O	N	O
Carte à bande	X		X		X		X		X		X		X		X	
Impression différée	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	X	X	X	X	X	X	X	X
Tri-fusion	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Histogrammes	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Tableaux de contrôle	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Indicateurs	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Comptours	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Mise à jour	(X)	(X)	(X)	(X)	X	X	X	X	X	X	X	X	X	X	X	X
Contrôles de structure			X	X			X	X			X	X			X	X
Contrôles validité et cohérence	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Messages									X	X	X	X	X	X	X	X
Amélioration des contrôles	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)	(X)
Polissage									X	X	X	X	X	X	X	X

X Fonction obligatoire (X) Fonction facultative Fonction absen

X-1- Organisation fonctionnelle de la configuration 1

(Voir l'organigramme ci-après)

Les fonctions obligatoires de la configuration 1 sont :

- le carte à bande,
- le tri-fusion,
- l'émission de compteurs,
- les contrôles de validité et de cohérence.

Nous retiendrons en outre les fonctions facultatives suivantes :

- histogrammes,
- tableaux de contrôle,
- production d'indicateurs qui entraîne la prise en compte de la fonction de mise à jour,
- amélioration des contrôles,
- impression en différé.

Sur l'organigramme les traitements se répartissent en six plages qui délimitent des phases de traitements homogènes du point de vue de leurs dates et de leurs fréquences d'intervention. Les phases sont numérotées dans l'ordre chronologique de leur intervention.

Phase I - Création du fichier directeur

Le fait qu'on veuille produire des indicateurs implique l'emploi d'un fichier directeur dont la création doit précéder l'apurement, voir le lancement de l'enquête.

On pourra créer le fichier directeur à partir d'un fichier de base préexistant à l'enquête et dont on tirera alors, en principe par sondage, la liste des unités statistiques à enquêter et même des questionnaires sur lesquels on préimprimera leurs données d'identification. Le fichier directeur sera alors une image de cette liste.

.../...

Si l'échantillon est constituée manuellement, que ce soit à partir de listes ou par un repérage sur le terrain, il faudra saisir les données d'identification sur cartes avant de constituer le fichier directeur. Il faudra alors le contrôler avec le plus grand soin car s'il contient des erreurs cela perturbera très sérieusement le fonctionnement du système.

Le fichier directeur sera stocké sur disque sous une méthode d'accès qui permette de le modifier sans le recopier. Il sera en effet de dimension assez restreinte (les variables directrices étant peu nombreuses) pour ne pas modiliser un espace disque trop important et pouvoir et pouvoir être interrogé aisément. De plus comme il va être modifié très fréquemment, le recopier à chaque fois sur un support séquentiel rendrait la gestion malaisée.

Phase II - L'apurement proprement dit

La phase II, dont l'objectif est de prendre en compte, contrôler et corriger les données d'enquête, est le coeur de la chaîne d'apurement. Elle est itérative dans l'hypothèse, où nous nous sommes placés, ou le contrôle manuel, la saisie et par conséquent le traitement sur ordinateur se font par lots. On y trouve successivement :

1 - le chargement des cartes sur support magnétique par un programme INTRO qui édite par ailleurs des compteurs de contrôle.

2 - Un tri sur des critères pertinents du point de vue des contrôles.

3 - Le bloc "CONTROLES ET REDRESSEMENTS" qui se décompose en un ou plusieurs programmes. Ici un seul devrait suffire ; en effet le fichier est saisi sur carte et il n'est pas structuré ce qui implique qu'il y a un seul type d'unité statistique dont les données sont en quantité suffisamment restreinte pour tenir sur une seule carte ; les contrôles et redressements ne doivent donc pas être très nonbreux. Le bloc utilise la fonction mise à jour puisque le fichier des données est apparié avec le fichier directeur. Il n'y a pas d'inconvénient important à mêler ici mise à jour et contrôles-redressements dans le même programme étant donné que nous nous trouvons dans le cas de deux fichiers non hiérarchisés.

En sortie le bloc fourni :

- des compteurs qui n'ont pas besoin d'être très détaillé vu l'existence du fichier directeur,

- le fichier des données d'enquête apuré,

- un fichier EVENEMENTS pour analyse ultérieure qu'on prévoiera de produire optionnellement pour les deux ou trois premiers lots de saisie seulement.

.../...

4 - Une fusion du fichier des données d'enquête apuré avec les lots apurés lors des précédents passages. Après apurement du dernier lot la sortie de fusion est le fichier définitif qui sera soumis à codification et tabulation.

Phase III - Histogrammes

On produira des histogrammes dès la saisie du 1er lot afin de détecter d'éventuelles erreurs systématiques qui ont pu se produire lors de l'enquête sur le terrain, du chiffrement ou de la saisie. On pourra ainsi prendre à temps les mesures correctives que ce soit au niveau du traitement manuel ou à celui du traitement automatique. Il suffira ensuite de produire à nouveau des histogrammes sur les deux ou trois lots suivants pour vérifier la prise en compte des corrections.

Phase IV - Tableaux de contrôle

On produira des tableaux de contrôle dans les mêmes conditions que les histogrammes pour affiner la détection des erreurs manuelles et détecter celles du programme de contrôles-redressements.

Phase V - Interrogation du fichier DIRECTEUR

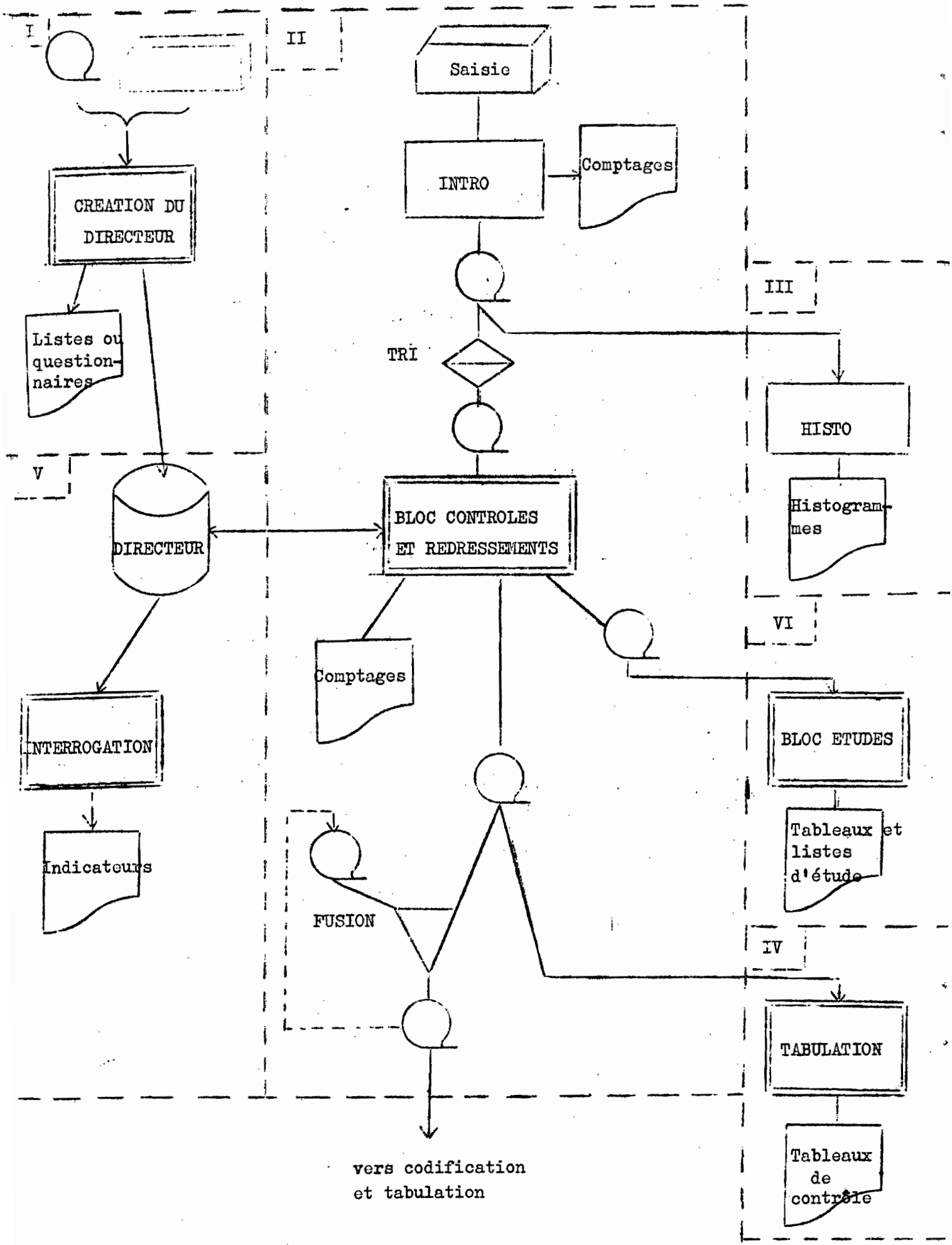
Elle interviendra à n'importe quel moment du processus d'apurement à l'initiative du statisticien et de l'analyste responsables de l'enquête afin de mesurer la qualité du fichier et l'avancement du travail. Ses résultats seront confrontés aux histogrammes et tableaux de contrôle dont ils pourront par ailleurs provoquer la production s'il révèlent des anomalies que la seule lecture des indicateurs ne permet pas de cerner complètement.

Phase VI - Etude des contrôles et redressements

Cette phase est pratiquement hors du champ de l'enquête en cours. L'étude minutieuse des effets des redressements qu'elle implique n'a d'intérêt que pour l'amélioration future d'une enquête permanente.

.../...

ORGANIGRAMME FONCTIONNEL DE LA CONFIGURATION 1



X-2- Organisation fonctionnelle de la configuration 16

Les fonctions obligatoires sont cette fois :

- l'impression en différé,
- le tri,
- l'émission de compteurs,
- la mise à jour,
- le contrôle de structure,
- les contrôles de validité et cohérence,
- l'émission de messages,
- le polissage.

Nous retiendrons comme fonctions facultatives :

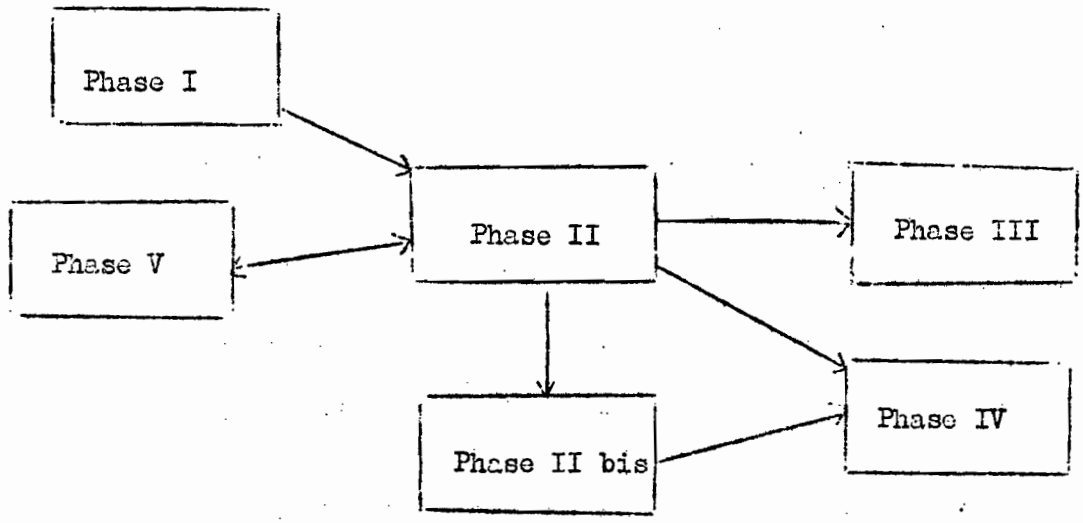
- les histogrammes,
- les tableaux de contrôle,
- la production d'indicateurs.

Nous conviendrons également que l'enquête utilise à la fois les redressements manuels et les redressements automatiques.

Le découpage en phase de la 1ère configuration reste valable ici à ceci près que l'apurement proprement dit (phase II) se décompose en deux parties : la première qui réalise les contrôles à redressement manuel est itérative ; la seconde qui réalise les redressements automatique et le polissage n'intervient une seule fois qu'après que le fichier a été apuré des erreurs à redressement manuel.

On peut schématiser l'organigramme fonctionnel comme ceci :

.../...



Il y a peu à ajouter à ce que nous avons déjà dit des phases I, III, IV et V sinon que la phase I aura pour fonction complémentaire la création du fichier des références externes qui est fonctionnellement très proche de celle du fichier directeur. Il est même souhaitable que ces deux fichiers soient confondus en un seul ce qui pourra se faire sans difficultés si les références externes sont limitées à quelques variables par unité statistique (1). On simplifiera ainsi la mise à jour qui n'aura à prendre en compte que trois fichiers au lieu de quatre. Quant aux autres elles resteront pratiquement identiques.

La phase II (apurement proprement dit) mérite par contre d'être analysée plus en détail car elle se complique notablement. L'organigramme ci-après en détaille le contenu. On y convient que mise à jour et contrôles sont répartis dans trois programmes qui traitent respectivement la mise à jour, les contrôles de structure, les contrôles de validité et de cohérence. Cette répartition devrait se rencontrer assez fréquemment. On trouve successivement les étapes suivantes :

.../...

(1) ce qui devrait être le cas le plus fréquent car, on l'a constaté, les variables externes pertinentes et fiables sont habituellement peu nombreuses d'une part, vouloir en intégrer un trop grand nombre risque de compliquer exagérément les contrôles d'autre part.

I - un tri sur identifiants et n° de lot du fichier en provenance de la saisie. La raison du tri sur identifiants est évidente. Le tri sur n° de lot est également nécessaire car il peut se faire, pour des raisons sans doute obscures, qu'entre deux passages on ait rédigé plus d'un bordereau de correction pour une même unité statistique.

II - Le programme de mise à jour dont les entrées sont :

- le fichier permanent issu du tour précédent (n-1),

- le fichier mouvement qui est le fichier saisi et trié,

- le fichier historique des mises à jour issu du tour précédent,

- une carte paramètre qui provoquera, si elle est active, le passage dans les contrôles de l'ensemble du fichier permanent (passage bilan),

- le fichier directeur,

En sortie on trouve :

- le fichier permanent du passage en cours (n),

- le fichier historique du passage en cours,

- le fichier des erreurs codées du passage en cours que le programme de mise à jour initialisé,

- un fichier "contrôle" qui dans le cas normal est un sous-ensemble du fichier permanent qui ne contient que les unités statistiques qui figuraient au fichier mouvement et pour lesquelles on a pas détecté d'erreurs de mise à jour ; il est en effet inopportun de soumettre à ces dernières à contrôle puisque la correction proposée ne s'est pas réalisée. Si la carte bilan est active, le fichier contrôle sera identique au fichier permanent aux unités statistiques en erreur de mise à jour près, bien entendu.

Une autre façon d'alimenter le fichier historique consisterait à fusionner, hors du programme de mise à jour, les fichiers mouvement n et historique n-1 ; il est d'ailleurs préférable de procéder ainsi car cela réduit le nombre des entrées-sorties du programme de mise à jour.

En cas de perte du fichier permanent ou d'erreur dans la gestion des programmes provoquant l'élimination de certaines unités statistiques on utilisera le fichier historique comme fichier de saisie pour reconstituer le permanent.

.../...

III - Le programme de contrôle de structure, qui reçoit en entrée le fichier contrôle issu de la mise à jour, consulte et renseigne le fichier directeur, alimente en sortie le fichier des erreurs codées et produit un fichier "contrôle 2" qui est une copie de "contrôle 1" élaguée des erreurs de structure car on a considéré qu'il était inutile de soumettre aux contrôles de validité et cohérence les unités statistiques incomplètes. Cette décision reste toutefois discutable.

IV - Le programme de contrôle de la validité et de la cohérence des données qui reçoit en entrée le fichier "contrôle 2", consulte le fichier directeur notamment pour y prélever les références externes qui n'interviennent qu'à ce niveau, le renseigne et alimente le fichier des erreurs codées. Le fichier "contrôle" disparait à cette étape. Il faut en effet se souvenir du fait qu'il n'est qu'une copie d'une partie du fichier permanent que les programmes de contrôle n'ont pas modifié puisque la correction est manuelle. Il est donc parfaitement redondant avec le fichier permanent.

Un problème semble toutefois se poser. Nous avons dit à propos du contrôle structure qu'il était souhaitable qu'il élimine les doubles sur identifiants du fichier permanent afin de ne pas perturber le bon fonctionnement de la mise à jour. Ce ne pourra ici être le cas puisque le fichier permanent n'est pas affecté par le contrôle de structure. On remarquera que si le programme de mise à jour, par lequel passent <u>toutes</u> les données saisies, est bien conforme aux spécifications que nous avons donné le cas ne pourra pas se produire puisqu'il considérera alors les multiplets sur identifiant comme des mises à jour successives et n'en retiendra que le dernier qui <u>seul</u> sera versé au fichier permanent. La prise en compte de l'ensemble de nos spécifications annule donc le problème des doubles. Dans le cas contraire il faudra toujours passer l'ensemble du fichier permanent dans les programmes de contrôle en reportant la carte bilan de la mise à jour dans chacun d'entre eux.

V - Un tri du fichier des erreurs codées qui permet de rapprocher les erreurs d'une même US détectées par les différents programmes et de les classer selon les critères de répartition des messages dans les ateliers de gestion. Dans notre exemple la 1ère fonction du tri est sans objet puisque chaque programme élimine pour les contrôles suivants les unités statistiques dans lesquelles il a détecté des erreurs. Il peut, dans d'autres cas, en être autrement.

VI - Le programme d'édition des messages qui, outre le fichier des erreurs codées trié reçoit en entrée le fichier permanent et un fichier disque qui contient les textes fixes des messages et éventuellement les spécifications des formats d'édition des données. Le recours au fichier de saisie n'est pas nécessaire puisqu'on a veillé à ce que les contrôles ne modifient en rien les données.

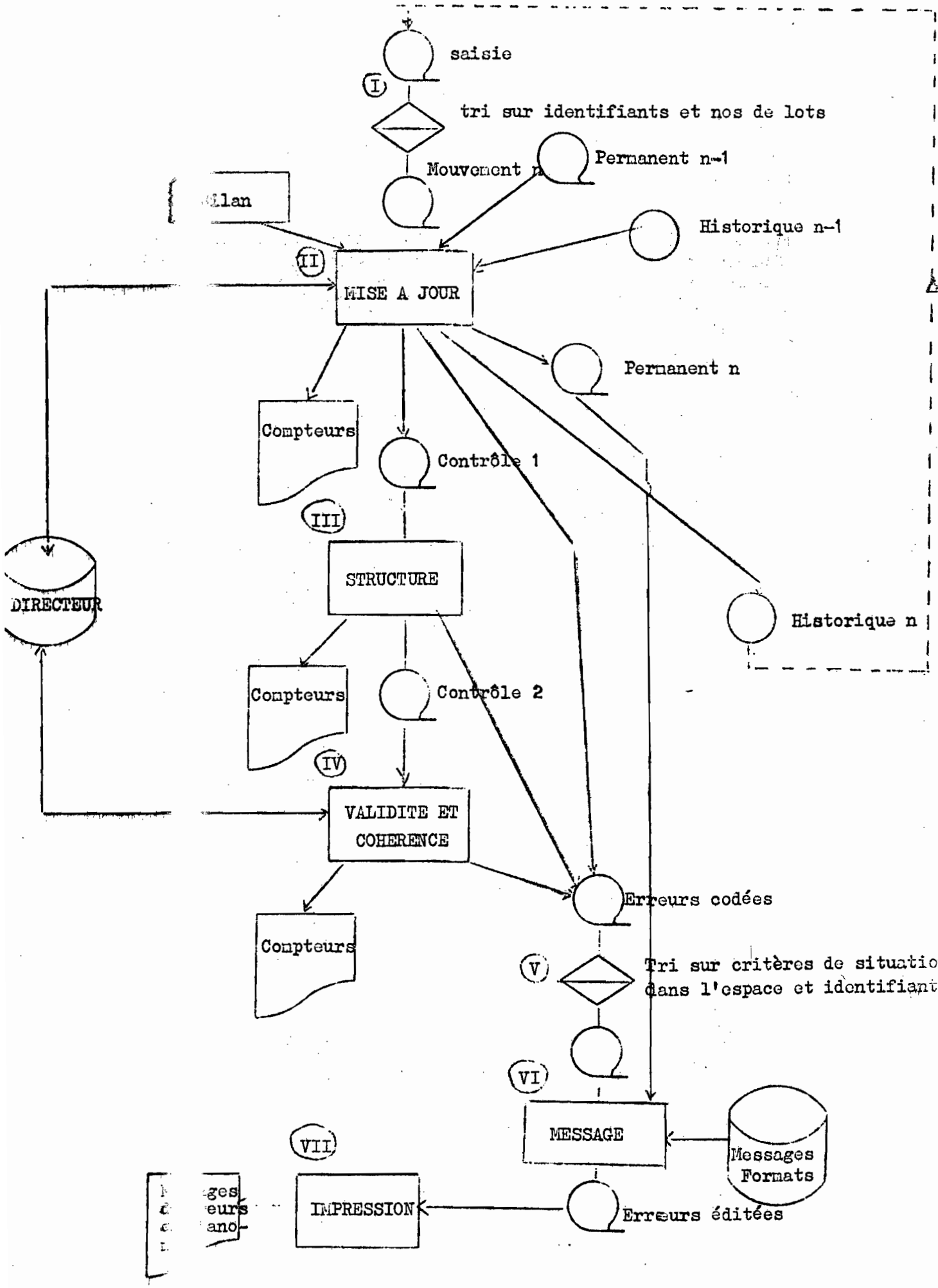
.../...

VII - Le programme MESSAGE produit un fichier des erreurs éditées qui sera imprimé en différé avant ventilation vers les ateliers de gestion.

Lorsque les interrogations du fichier directeur indiqueront que le fichier permanent est complet (ou presque) et qu'il n'y subsiste plus d'erreurs à correction manuelle on pourra le soumettre à la phase II bis d'apurement automatique et polissage qui pourront être réunis dans un seul programme. On disposera alors d'un fichier exploitable.

.../...

C: IGRAMME FONCTIONNEL DE L'APUREMENT - CONFIGURATION 16



XI - CONCLUSION

La chaîne d'apurement telle que nous venons de la présenter traite, nous l'avons dit en introduction, par lots sur un ordinateur de moyenne ou grosse configuration des fichiers de données stockés sur des supports à accès séquentiel. Nous pensons, dans ce cadre avoir fait, à peu près, le tour du problème. On peut, certe, imaginer des solutions différentes en ce qui concerne la résolution de certaines fonctions et l'agencement des chaînes de traitement. Il reste malgré tout que les fonctions utilisées feront partie de celles que nous avons décrites et que leur enchaînement sera celui que nous proposons. Si le lecteur en est maintenant convaincu et s'il en a déduit que la réalisation d'une chaîne d'apurement d'enquête pose une série de problèmes bien caractérisés auxquels on apporte des solutions types dans certains cas bien standardisés, nous avons atteint notre objectif.

Il reste à se demander ce qui peut résulter d'hypothèses de travail différents. Nous en examinons trois qui sont :

- l'utilisation d'ordinateurs de petite configuration,
- le stockage des données d'enquête sur des supports à accès direct,
- le contrôle et la correction des données en ligne.

XI-1- Ordinateurs de petite configuration

L'organisation du système ne se trouvera en rien modifiée car elle est complètement induite pour le traitement par lots et l'utilisation de supports séquentiels.

Les fonctions d'entrées-sorties (carte-à-bande et impression différée) deviendront éventuellement inutiles si l'ordinateur ne fait que de la monoprogrammation.

Il faudra peut être décomposer les traitements en un plus grand nombre de programmes, notamment au niveau du bloc de contrôle, pour qu'ils tiennent dans un espace mémoire restreint. Encore que la segmentation des programmes (ou overlay) permette de pallier cette difficulté.

.../...

Il pourra également se révèler nécessaire, dans le cas de [c]onfiguration 16 par exemple, d'aménager le programme de mise à jour [d]elle sorte qu'il utilise poins de phériphériques d'entrée-sortie [d]e notre exemple il utilise 7 ou 5 lecteurs de bandes magnétiques [s]uivant qu'il intègre ou non l'alimentation du fichier historique). Ce [t]ravail sera être fait en réunissant, en sortie, le fichier permanent, le fi[c]hier "contrôle" et le fichier des erreurs codées sur un même support. [L]es programmes suivants devront alors être aménagés de manière à ne [t]raiter sur le fichier qui en résulte que les informations qui les con[c]ernent.

XI-2- Supports à accès direct

Le principal obstacle au stockage de toutes les données d'en[q]uêtes sur disques reste, aujourd'hui encore, la cherté de ce type de sup[p]ort. L'argument a toutefois perdu beaucoup de son importance depuis quel[q]ues années et il est probable qu'il deviendra de plus en plus mince.

Le stockage sur disque présente par contre l'intérêt majeur [d]'rendre possible des passages très fréquents (hebdomadaires ou quoti[d]iens par exemple) de la chaîne d'apurement lorsqu'on utilise les redres[s]ements manuels. En effet l'inconvénient qu'il y a à dérouler entière[m]ent des fichiers séquentiels, de plus en plus volumineux au fur et à [m]esure des passages, pour prendre en compte un volume marginal d'infor[m]ations, d'autant plus restreint que les passages sont plus fréquents [n]'existe plus. Par contre il est du plus haut intérêt de restituer aux [e]nquêteurs dans le délai le plus court possible les résultats des con[t]rôles et corrections qu'ils ont effectués.

Les données directrices (ou indicateurs), les références ex[t]ernes et les données d'enquêtes pourront alors être réunies dans un fi[c]hier unique à accès direct. Le fichier historique, et éventuellement [l]e fichier évènements, restent sur des supports à accès séquentiel car [i]l n'y a pas de raison particulière de les consulter directement.

La mise à jour et les contrôles peuvent être intégrés dans [u]n seul programme (l'accès direct simplifie ou annule purement et sim[p]lement les problèmes de la mise à jour en séquentiel) où ils interven[d]ront dans le même ordre que précédemment. Simplement au lieu de soumettre [l]'ensemble des unités statistiques concernées par un lot de mise à jour [d]'abord à la mise à jour, puis aux contrôles de structure, puis aux con[t]rôles de validité et cohérence on soumettra chacune d'elle à ces trois [o]pérations successivement. Ce programme, qui doit être modulaire, risque [d]'être très volumineux. On devra donc probablement le segmenter.

L'édition des messages elle pourra être intégrée au programme [d]e contrôle ou en rester séparée. Il ne semble pas qu'une solution pré[s]ente d'intérêt particulier par rapport à l'autre. Quant à la production [d]'histogrammes et de tableau de contrôle, de même que les corrections au[t]omatiques et le polissage ils restent identiques à ce qu'ils étaient aupa[r]avant.

.../...

VI-3- Contrôles et corrections en ligne

Il faut ici distinguer plusieurs cas selon qu'on réalise tout ou partie des corrections en ligne.

1er cas

Il est en fait inclus dans ce qui précède. C'est celui où la saisie se fait sur support magnétique à l'aide d'un matériel multi- ou mono-claviers muni d'une petite unité centrale. On peut sur un tel matériel programmer un certain nombre de contrôles de structure de validité et de cohérence simple des données. On sera arrêté certe par la faible configuration mémoire du matériel, mais surtout par le fait que le travail est réalisé par un personnel spécialisé en saisie, soumis à des contraintes de rendement, et qui n'a aucune connaissance particulière de l'enquête traitée. Il faudra donc liniter les contrôles à la détection des erreurs de saisie. Cela permettra d'alléger le volume des erreurs, notamment de structure, détectées par la chaîne d'apurement mais celle-ci ne sera en rien modifiée.

2ème cas

La saisie est faite par les gestionnaires de l'enquête eux-mêmes sur des micro-ordinateurs (1) portables qu'on peut même envisager d'emmener sur le terrain pour faire la saisie auprès de l'enquêté ! Le programme de contrôle sur micro-ordinateurs enchaînera, encore une fois, les contrôles dans le même ordre. Si à ce stade contrôles et corrections ont été poussés assez loin on pourra, par exemple, avoir sur gros ordinateur une chaîne à corrections entièrement automatiques du type de la configuration 1 présentée dans le paragraphe X dans laquelle le fichier directeur aura pour principale fonction de contrôler l'exhaustivité du fichier ; les fonctions "histogrammes" et "tableaux de contrôle" permettront de détecter les biais introduits dans le redressement manuel par tel ou tel gestionnaire ; l'étude des évènements perdra beaucoup de son intérêt car les corrections automatiques ne portent alors que sur des variables secondaires.

.../...

(1) D'un point de vue matériel le micro-ordinateur ne se distingue pas d'un appareil de saisie monoclavier si ce n'est par le fait qu'il peut être plus compact et plus léger. Il est seulement muni d'un logiciel plus évolué qui améliore les possibilités de programmation.

3ème cas

Tous les contrôles et corrections se font en ligne sur gros ordinateur par les gestionnaires de l'enquête à partir de terminaux.

On peut se demander si cette solution est bien opportune. La saisie doit en effet être faite par un personnel hautement qualifié (car la correction des erreurs demande une parfaite connaissance du sujet traité) qui risque de mal accepter la perspective de passer une part importante de son temps rivé à un clavier de terminal. De plus la correction des erreurs n'est pas toujours immédiate, elle requièrt souvent qu'on relise une partie du questionnaire, qu'on consulte des sources externes, voire qu'on reprenne contact avec l'enquêté. Il peut donc s'écouler plusieurs heures ou plusieurs jours entre le moment ou l'erreur est détectée et celui où on dispose des éléments nécessaires à la correction qui devra être différée. Il sera souhaitable de sortir sur papier un message d'erreur, répondant aux normes que nous avons définies, pour garder une trace et ne pas oublier le problème. Cette solution ne semble donc présenter d'intérêt que pour les enquêtes rapides, dont le volume de saisie par unité statistique est peu important, dont les règles de correction des erreurs sont simples et immédiates et qui sont, de plus, répétitives car les charges d'analyse-programmation et de formation du personnel de saisie sont très notablement supérieures à ceux d'un traitement par lots. On sort donc en principe du champ des enquêtes démographiques qui présentent rarement ces caractéristiques.

L'architecture du système informatique est bien évidemment
te de celle du système en traitement par lots. Il n'est pas dans notre propos de l'analyser. Elle sera d'ailleurs sous-tendue par des logiciels généraux de saisie en ligne fournis par les constructeurs ou des sociétés de service. On y retrouvera toutefois, comme dans le cas précédent, les principales fonctions logiques de l'apurement toujours identiques quant à leurs spécifications fonctionnelles.

../...

Chapitre 9

La tabulation

Rédacteurs : M. EURIAT,
G. GRENIER,

Etat de la rédaction : définitive,

Plan du Chapitre :

- Choix des tableaux,
- Les relations statisticien - informaticien,
- Organisation logique d'un programme de tabulation,
- Evaluation d'un logiciel général de tabulation,
- Gestion des tableaux.

M. EURIAT
G. GRENIER

TABULATION

INTRODUCTION

Le résultat principal que l'on attend d'une enquête est bien évidemment la fourniture d'un certain nombre de tableaux statistiques qui résumeront l'information recueillie. Ces tableaux seront choisis et dessinés par le statisticien et réalisés selon un processus particulier par l'informaticien.

Nous n'aborderons pas ici le problème du contenu et de la signification statistique de ces tableaux mais nous nous intéresserons à leur "dessin" et à leur faisabilité par rapport aux données collectées et au fichier informatique.

Nous évoquerons donc les problèmes posés lors de la confection de la liste de tableaux, puis ceux de la communication entre le statisticien et l'informaticien grâce au "langage" de tabulation commun.

Dans le paragraphe 3 nous indiquerons l'organisation logique d'un programme de tabulation pour l'informaticien. Enfin nous proposerons une grille d'évaluation pour les logiciels généraux et quelques conseils pour la gestion des tableaux.

I - LE CHOIX DES TABLEAUX

En pratique le statisticien dessine rarement les tableaux qu'il veut obtenir avant de commencer son enquête.

Et les tableaux sont souvent conçus indépendamment du questionnaire et à la fin du processus informatique de dépouillement. Cette façon de procéder présente des inconvénients et peut conduire à de graves ennuis: nous le montrerons plus loin sur un exemple. Il conviendrait donc plutôt de partir du dessin des tableaux, même approximatif, pour réaliser le questionnaire afin d'y poser les questions ad hoc ; la saisie, le contrôle et l'organisation des données seraient alors guidés par la forme des tableaux à produire. Nous voyons donc une fois de plus que l'enquête n'est pas un processus linéaire mais que les différentes phases qui la composent sont en interaction.

Avant de passer en revue quelques écueils à éviter dans le choix des tableaux, nous croyons utile de rappeler ici quelques définitions concernant les tableaux statistiques qui seront utilisés par la suite, ces définitions étant souvent peu familières au non-statisticien.

1) TABLEAUX A DOUBLE ENTREE

1.1) Tableaux d'effectifs.

Nous considérons la population formée par l'ensemble des N unités statistiques de niveau 1, ici les ménages. Chaque ménage est décrit simultanément par des critères par exemple Région R et profession du chef de ménage P.

soient $R1, R2, \ldots\ldots\ldots\ldots R_k$ les k modalités du critère R

$P1, P2, \ldots\ldots\ldots\ldots P_m$ les m modalités du critère P

Ces deux critères vont définir un tableau où les modalités de R seront rangées en colonne par exemple et celles de P en ligne.

Soit N_{ij} le nombre de ménages présentant à la fois la modalité R_i et la modalité P_j. Les modalités de R sont incompatibles entre elles et tout ménage est caractérisé par une modalité. Il en est de même de P.

La somme des effectifs N_{ij} est donc égale à la taille de la population :

$$\sum_{i=1}^{k} \sum_{j=1}^{m} N_{ij} = N$$

On rangera dans le tableau qui comporte k colonnes et m lignes à l'intersection de **la ième ligne et de la jème colonne de la quantité Nij.**

On complète souvent le tableau par des totaux verticaux ou horizontaux. A la modalité R_i on associera la quantité :

$$N_{i\bullet} = \sum_{j=1}^{m} N_i$$

qui est le nombre de ménages présentant la modalité R_i

Il en est de même pour :

$$N_{\bullet j} = \sum_{i=1}^{k} N_{ij}$$

De plus :

$$N = \sum_{i=1}^{k} \sum_{j=1}^{m} N_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{k} N_{ij} = \sum_{j=1}^{m} N_{\bullet j} = \sum_{i=1}^{k} N_{i\bullet}$$

On a désigné par un • la totalisation suivant l'indice i ou j

Répartition des ménages selon la Région et la Profession du chef de ménage.

modalités du critère R / modalités du critère P	R1	R2	R3		Rk	TOTAUX HORIZONTAUX
P1	N_{11}	N_{12}	N_{13}		N_{1k}	$N_{1\bullet}$
P_1	N_{21}	N_{22}	N_{23}		N_{2k}	$N_{2\bullet}$
P_3	N_{31}	N_{32}	N_{33}		N_{3k}	$N_{3\bullet}$
P_m	N_{m1}	N_{m2}	N_{m3}		N_{mk}	$N_{m\bullet}$
TOTAUX VERTICAUX	$N_{\bullet 1}$	$N_{\bullet 2}$	$N_{\bullet 3}$		$N_{\bullet k}$	N

1.2 - Tableau de pourcentages ou de fréquences.

Pourcentage ligne :

Il s'agit d'un tableau identique au précédent; cependant on remplace à l'intersection de la ligne i et de la colonne j la quantité N_{ij} par $(N_{ij}/N_{i\bullet})$x100. On obtient ainsi la distribution du critère R pour une valeur fixée de la modalité de P.

Pourcentage colonne

On remplace N_{ij} par $(N_{ij}/N_{\bullet j})$ x 100.
Les totaux verticaux sont tous égaux à 100.
Les totaux horizontaux à $(N_{\bullet j}/N)$ x 100.

Pourcentage total :

On remplace N_{ij} par (N_{ij}/N)x100

2) Tableaux de quantités ou de moyennes

Nous avons vu qu'en plus des critères chaque ménage était caractérisé par des quantités Q, par exemple le revenu.

Si nous voulons connaître le revenu des ménages selon la région et la profession nous construirons un tableau identique au précédent. Cependant nous cumulerons les revenus de tous les ménages appartenant à la région j et à la profession i. Soit Q_{ij} le revenu, Q_{ij} remplaçant N_{ij} dans le tableau on définit de même :

$$Q_{i\cdot} = \sum_{j=1}^{k} Q_{ij} \qquad \text{et} \qquad Q_{\cdot j} = \sum_{i=1}^{m} Q_{ij} \qquad Q_{\cdot\cdot} = \sum_{i=1}^{m} Q_{i\cdot}$$

$Q_{i\cdot}$ est donc le revenu du ménage de la profession i

$Q_{\cdot j}$ celui des ménages de la région j

Si nous voulons connaître le revenu moyen des ménages selon la profession et la région,

il nous faudra remplacer dans le tableau Q_{ij} par (Q_{ij}/N_{ij}). Il est clair que si N_{ij} est nul Q_{ij} l'est aussi, on mettra donc un zéro à l'intersection de la ième ligne et de la jème colonne. Du fait de la non additivité des moyennes les totaux horizontaux et verticaux deviennent $Q_{i\cdot}/N_{i\cdot}$ et $Q_{\cdot j}/N_{\cdot j}$, la moyenne de la population entière $Q_{\cdot\cdot}/N$

3) Tableaux faits sur plusieurs niveaux hiérarchiques.

En reprenant l'exemple du ch 7 au niveau 2 nous avons les individus du ménage.

Nous pouvons reprendre les trois tableaux précédents en remplaçant N_{ij} par le nombre d'individus des ménages de la Région j et dont la profession du chef de ménage est i

Un autre type de tableau serait par exemple la répartition des individus selon la région et la tranche d'âge. Chaque individu est caractérisé par le critère tranche d'âge A.

Le tableau résulte du croisement des critères R et A. L'élément N_{ij} est le nombre de personnes de la région i dont la tranche d'âge est j.

Enfin le tableau revenu moyen par personne selon la région et la tranche d'âge :

à l'intersection de la ligne i (région i) et de la tranche d'âge j on trouvera la quantité :

$$Q_{ij}=Q_{i.}/N'_{ij}$$

$Q_{i.}$ revenu des ménages de la région i ; N'_{ij} nombre de personne de la région i et de la tranche d'âge j

Les tableaux faits sur plusieurs niveaux hiérarchiques et en particulier les tableaux de moyennes peuvent prêter à ambiguïté. Il importe que l'informaticien se fasse préciser avec le plus grand soin par le statisticien tous les éléments du calcul de la fraction.

4) Appartenances et filtres

-Le critère d'appartenance permet de sélectionner une sous-population grâce à une condition logique sur un ou plusieurs critères (C=1 et D=2)...etc

Exemple : Répartition des ménages urbains selon la région et la profession.

On sélectionne la population des ménages urbains puis on procède comme au 1) pour obtenir la répartition par région et profession.

- Le filtre permet de ne cumuler une quantité dans une case du tableau que si une expression logique associée à la case du tableau est vérifiée

Exemple : Revenu des ménages gagnant moins de x francs selon région et profession

On retrouve le tableau définit en 2) cependant on ne cumulera dans ce cas que les revenus inférieurs à x francs.

Signalons là aussi quelques difficultés ainsi le tableau revenu moyen des ménages gagnant moins de x francs selon région et profession. Il faudra dans chaque case diviser le revenu calculé précédemment par le nombre de ménages de la région i de la profession j et dont le revenu est inférieur à x francs.

• Tableaux statistiques à multiples entrées.

C'est une généralisation du cas précédent. On croise alors plusieurs critères. Exemple : dans le cas de trois critères l'élément du tableau devient N_{ijk}, nombre de ménages appartenant à la modalité i du premier critère, j du second et k du troisième.

Un cas assez fréquent est celui du critère page.

Exemple : Répartition des ménages selon le type de résidence, la région et la profession,

Ce rappel effectué venons en à quelques points que le statisticien devra garder présent à l'esprit quand il établira sa liste de tableaux.

- Les tableaux aberrants. Ventiler mille unités statistiques dans un tableau de dix mille cases et une opération formellement réalisable mais de peu d'intérêt statistique. On vérifiera donc si les critères que l'on croise n'ont pas de listes de modalités trop longues par rapport à la taille de la population concernée.

- Il convient aussi de ne pas abuser du croisement de plus de 3 critères dans un tableau, la lecture et l'interprétation de tels tableaux devenant rapidement difficiles.

- On évitera aussi de demander le croisement systématique de tous les critères 2 à 2 sous prétexte que "l'informatique peut le faire". Si l'on ne peut sélectionner a priori les croisements intéressants on pourra avec profit faire calculer en même temps que le tableau un indicateur de dépendance des critères. On peut citer dans le cas des tableaux de contingence (I,J) le lien entre I et J noté L(I,J).

$$L(I,J) = \sum_i \sum_j fi.\ f.j \left[\frac{fij - fi.\ f.j}{fi.\ f.j} \right]^2 = \sum_{i,j} Cij$$

où

$fij = N_{ij}/N$ $\quad fi. = N_{i.}/N$ et $\quad f.j = N_{.j}/N$

On montre que ce lien est égal à un coefficient près à l'information mutuelle de I et de J. On voit d'autre part que $L(I,J)=0 \Leftrightarrow$ I et J indépendants.

Il sera intéressant de rechercher parmi tous les couples de critères (I_k, I_l) ceux pour qui le croisement apporte le plus d'information : ce sont les tableaux pour lesquels le lien $L(I_k, I_l)$ est le plus fort. Cela fournit donc une méthode a posteriori pour sélectionner les tableaux les plus significatifs.

Enfin pour un tableau donné (I,J) on peut classer les Cij par ordre croissant, on détermine ainsi les cases du tableau qui apportent le plus d'information.

Pour plus de détails sur cette méthode nous renvoyons le lecteur à l'article de M. VOLLE paru dans Economie et Statistique INSEE PARIS Janvier 1974.

- Les contraintes informatiques.

Pour rationaliser le processus de dépouillement l'informaticien a opéré des choix dans la saisie des données puis dans la structuration et l'apurement de son fichier (cf les chapitres précédents). Ces choix peuvent rendre difficile voire impossible la réalisation de certains tableaux. Citons une des principales limitations qui résulte de la hiérarchie des fichiers séquentiels telle qu'explicitée au chapitre 7 : il n'est pas possible de croiser des critères qui ne sont pas en filiation.

L'informaticien devra donc insister auprès du statisticien pour que ce dernier lui fournisse assez tôt la liste des tableaux. Un exemple fera sentir l'influence de la forme des talbeaux à sortir sur le hiérarchisation du fichier.

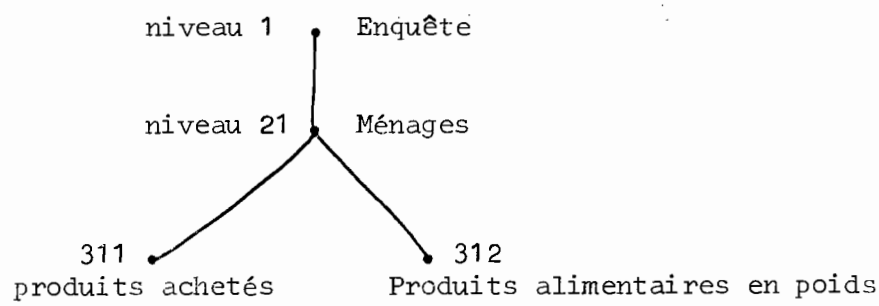
Supposons une enquête Budget-Consommation-Nutrition auprès des ménages. Nous avons 3 types de questionnaires.

1 - Questionnaire comportant différentes questions sur le caractéristiques du ménage.

2 - Questionnaire sur les produits achetés par le ménage en vue de la consommation.

3 - Questionnaire sur les produits alimentaires (en poids) ce questionnaire est effectué selon une technique différente de celle du questionnaire 2 car il va servir de base à l'analyse de la nutrition. Les produits alimentaires achetés seront dans le questionnaire 2 en valeur, les produits autoconsommés ou achetés seront dans le questionnaire 3 avec leurs poids.

Il peut sembler naturel d'organiser le fichier de la façon suivante :



On saisira les données et on les apurera sous cette forme. Et on pourrait les organiser ainsi grâce à un logiciel de type LEDA.

Mais supposons que l'on veuille fabriquer le tableau suivant. Répartition de la consommation des ménages selon les produits :

N° de produit	achats	autoconsommation valorisée	total
111	--	---	--
--	--	---	--

L'organisation hiérarchique est incompatible avec la fabrication du tableau.

Il faudra revenir au fichier de départ, le trier selon les clés de tri (ménage, produits, type de questionnaire) puis créer pour chaque ménage autant d'enregistrement que de produits consommés. L'enregistrement produit comprenant les 2 parties achats, et produits alimentaires

N° de Pdt	Achats	Pdts alim

une des 2 parties pouvant être vide.

La nouvelle structure hiérarchique devient :

niveau1 . Enquête

niveau 21 . Ménage

niveau 311 . Produit

Nous voyons sur cet exemple que la structure logique du fichier ne résulte pas automatiquement de l'organisation des données dans le questionnaire, ou le document de saisie, mais de la forme de l'état de sortie.

- Le statisticien devra s'assurer qu'il possède bien la liste exacte et à jour des différentes variables contenues dans le fichier avec la liste des modalités des codes, ainsi que des modifications apportées aux données par l'informaticien (redressement). Il n'utilisera alors pour définir ses tableaux que des éléments figurant réellement dans le fichier informatique.

- Enfin dernier type de contrainte, les limitations de la machine. Certains tableaux peuvent requérir une place trop importante en mémoire centrale pour le matériel utilisé. Il convient alors de modifier l'exploitation pour les réaliser. Exemple : passage d'une ventilation fermée à une ventilation ouverte (ces termes seront définis au §3). Signalons aussi les contraintes liées à la mise en page, aux libellés et aux formats (cf aussi ch.10).

- Organisation de la production des tableaux.

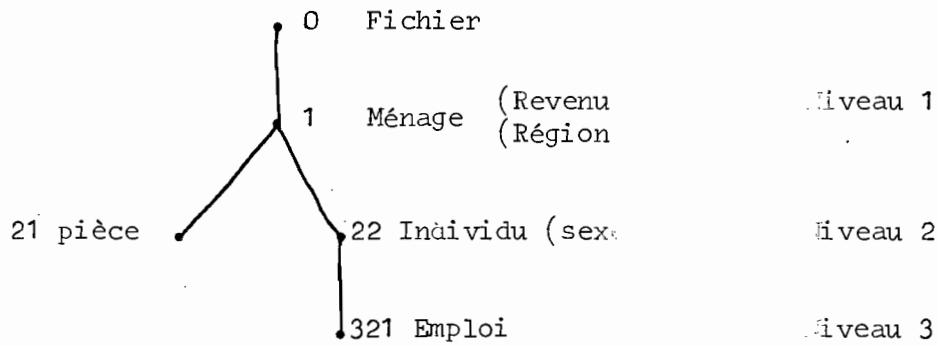
Il contient d'organiser la production en 3 temps. D'abord les tableaux urgents et prioritaires en petit nombre, qui permettent de faire un premier point sur la qualité de l'enquête. Puis les tableaux intermédiaires ou de travail qui peuvent servir à définir de nouvelles variables (confection de tranches pour des variables quantitatives, regroupement des nomenclatures...). Enfin les tableaux définitifs qui ne seront commandés qu'après vérification soigneuse des deux premières séries. Ce système permet de tester la bonne compréhension entre le statisticien et l'informaticien, de déceler les erreurs ou les incohérences de fichiers. Les modifications éventuelles peuvent alors être apportées car tous les "crédits informatiques" n'ont pas encore été dépensés. Une programmation des moyens et du financement doit donc être prévue dans ce sens au début de l'enquête.

On peut signaler une autre solution qui consisterait en le stockage sur support informatisé du fichier hiérarchisé avec un "dictionnaire" : on évite la production d'un grand nombre de tableaux en ayant recours a une production à la demande. On peut aussi stocker des tableaux eux-mêmes d'une utilisation ultérieure voir à ce sujet le §5.

II - LES RELATIONS STATISTICIEN-INF ICI

Pour que la demande du stat ien précise et correctement interprétée par l'informaticien il ent l y ait entre eux un langa- ge commun. Ce langage de tabulation per d'échanger des documents et de communiquer sans ambiguïté. N rése s la terminologie utilisée par le système LEDA en service à l' com n exemple de langage de tabulation.

Nous supposons les données o isée érarchiquement et reprenons l'exemple du chapitre 7 pour illustr otre pos :



Rappelons que chaque niveau cons é de plusieurs unités statistiques.

Chaque unité statistique est acté e par des critères (ou code ou variable qualitative) et des ntit ou variable quantitative).

- Un critère permet d'associer à cha U.S. nombre entier (modalité du critère).

La liste des modalités possi pou critère donné est connue d'avance ou à la fin de l'enquête.

Ex. : le sexe pour l'U.S. in du l associée (1,2)

- Une quantité permet d'associer à c e U. n nombre réel donné dont la valeur est inconnue a priori

Ex. : revenu du ménage pour S. m e

Certains critères peuvent de r de antités et réciproquement.

Eléments nécessaires à la définition d'un tableau

Pour définir un tableau il faudra se fixer le niveau hiérarchique c'est-à-dire la population concernée.

Le tableau étant caractérisé par le croisement de 2 ou plusieurs critères ou par 1 critère, la population à ventiler correspond au niveau hiérarchique le plus élevé.

Ex. : au tableau région x sexe est associée la population des individus niveau 2

On peut vouloir spécifier une sous-population qui sera seule concernée par le tableau en question.

On définit cette sous-population par des critères d'appartenance liés entre eux par une condition logique et prenant des modalités définies

Ex. : population : les ménages

Sous-population définie par les ménages de la région 1 et de la catégorie sociale 33.

Critère d'appartenance : région et CS

Valeur prise par ces critères (région = 1 et CS = 33).

Cumuls ou quantités à cumuler

Chaque case du tableau ainsi défini est remplie par une ou plusieurs quantités ou sous-case à cumuler pour les U.S. de la population.

Filtres sur les cumuls

On peut imposer des conditions sur les sous-cases pour que ces cumuls aient lieu.

Ex. : on cumule dans chaque case d'un tableau de poids consommé d'un produit et sa valeur. On définit un filtre en ne cumulant que les produits dont le code est compris entre 1 et 3.

Calculs entre cumuls (ou entre sous-cases)

On peut vouloir opérer des opérations entre les cumuls d'une même case pour définir une nouvelle quantité. On se fixe alors des règles de calculs. La nouvelle quantité sera calculée après le remplissage des cases.

Ex. : on peut définir un prix unitaire en faisant le rapport des quantités valeur/poids.

Le tableau est alo peu près défini. Il reste à fixer la mise en page du tableau :

- le titre en clai le n° du tableau éventuellement
- la place des diff ts critères en ligne, en colonnes ou en page
- l'existence ou nc libellés en clair pour des modalités
- le format d'impre n des nombres.

Il reste à préciser l'on veut des pourcentages (ligne, colonne ou sur le total) ou des nom bruts, des sous-totaux (ligne, colonne), ou des récapitulations sur les miers chiffres d'un code analytique.

Ces éléments spécifi on peut arriver à une présentation standardisée des demandes de table et cela que l'on utilise un logiciel général ou un programme particulier.

Dans le cas de l'uti ation d'un logiciel général, il n'y a pas programmation. Il conviendra confier au statisticien dans la mesure du possible, l'écriture des ins tions nécessaires au logiciel.

On peut concevoir al des bordereaux de demandes de tableaux standardisés. Un exemple de bordereau dans le cas du logiciel LEDA est fourni en annexe.

Si l'on ne dispose p 'un logiciel général la fourniture pour chaque tableau de la liste d léments du tableau simplifiera le travail du statisticien et de l'info icien.

Pour les tableaux do e dessin est assez compliqué on pourra faire une maquette grandeur re en s'aidant d'une grille d'imprimante.

La collaboration entr e statisticien et l'informaticien s'avère donc nécessaire dans tout le essus de tabulation.

L E D A phase HELENE

Si HLN 91-1

IMPRIME DE DEMANDE DE TABLEAU

Commentaire :

Feuille n°

/*

*/

(1) P A R A M E T R E S

(2) T I T R E

col. 8

(3) V E N T I L A T I O N N U M E R O

(4) A P P A R T E N A N C E S

#

(5) page C R I T E R E S

ligne

#/

colonne

#/

(6) F I L T R E S

#

#

#

(7) C U M U L S

#

#

#

(8) C A L C U L S

#

#

#

(9) E N - T E T E S

#

#

#

(10) F O R M A T S

#

#

#

Comment remplir l'Imprimé de demande de tableau ? (voir SI LEDA 80)

Généralités : pour chaque rubrique la dernière ligne remplie doit être terminée par un (;)

1 - PARAMETRES : Facultatif. Ecrire la liste des paramètres séparés par une virgule (,).

2 - TITRE : Donner le libellé du titre du tableau. Tous les caractères sont permis à l'exclusion du caractère ('). Le nombre de caractères imprimés par ligne est 96. (Cette séparation est indiquée par une surcharge sur l'imprimé). Deux lignes au maximum sont prévues sur l'imprimé.

3 - NOM DE LA VENTILATION : Donner un nom symbolique pour identifier le tableau. Pour tous les tableaux d'un même fichier ils doivent tous être différents.

VENTILATION NUMERO : Les quatre caractères soulignés représentent le numéro du tableau qui figurera à l'édition dans la ligne "Label".

4 - APPARTENANCES : Facultatif. Indiquer les expressions logiques qui définissent la sous-population à ventiler.

Ex : APPARTENANCES (CSP = '00') | (CSP = '99')
#SEXE = '1' ;

5 - CRITERES : Chaque critère de ventilation est séparé par un (#). Sur la première ligne donner les critères "page", sur la deuxième ligne les critères" ligne" sur la troisième ligne les critéres "colonne".

6 - FILTRES : Facultatif. Donner les différentes expressions des filtres relatifs aux sous-cases. Chaque expression doit être séparées par un (#). Utiliser à cet effet les cases réservées (en surcharge).

7 - CUMULS : Donner les différentes expressions des cumuls relatifs aux sous-cases. Pour l'écriture voir (6). Attention à la correspondance FILTRES-CUMULS.

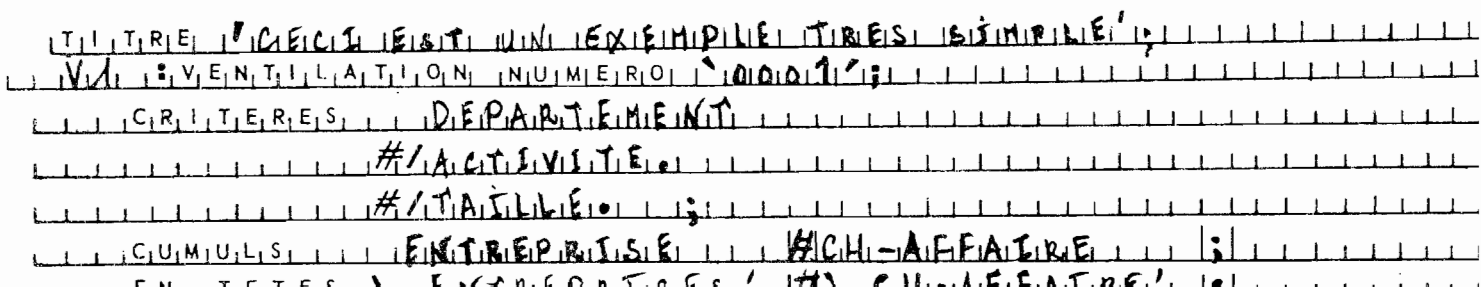
8 - CALCULS : Facultatif. Ces calculs font intervenir les sous-cases de l'instruction CUMULS : SCS (n) = neme sous case. Pour l'écriture voir (7).

9 - EN-TETES : Facultatif : Permet de donner une en-tête à chaque sous-case résultante. Ces en-têtes sont relatives aux sous-cases calculs, sinon s'il n'y a pas de calculs, elles sont relatives aux sous-cases cumuls. Chaque en-tête est encadrée par deux ('), et séparée des autres par un (//).

10 - FORMATS : Facultatif. Définit le format d'édition des sous-cases résultantes des calculs ou des cumuls s'il n'y a pas de calculs. Ecriture : voir (9).

NOTA : Si vous désirez utiliser une recodification, une liste, des options d'édition, un supercode, décrivez-lez sur une feuille à part. Demême si vous désirez des intitulés pour un code, décrivez-les sur une feuille à part et indiquez l'option INTITULE pour le code dans l'instruction CRITERES.

EXEMPLE :



1099.73.D

§ 3. ORGANISATION D'UN PROGRAMME DE TABULATION.

Nous présentons quelques remarques sur la façon dont on peut écrire un programme, ou un ensemble de programmes permettant d'obtenir des tableaux à partir d'un fichier de questionnaires. Ces remarques s'appliquent aussi bien à un programme spécifique à une enquête qu'à un programme général de tabulation : elles peuvent donc servir :

- soit à concevoir une chaîne de programmes de tabulation,
- soit dans le cas de l'utilisation d'un programme général, à apprécier la manière dont il fonctionne, ce qui peut influer sur la manière de l'utiliser.

Le programme le plus simple consisterait à prévoir un certain nombre d'emplacements de mémoire, correspondant chacun à un des nombres qui figureront dans le tableau à réaliser. On lit alors le fichier d'enquête enregistrement par enregistrement, et on cumule les données lues, lorque les conditions d'appartenance et les filtre sont satisfaits, dans les bons emplacements. Bien que cette méthode soit au fond celle qui sera toujours employée, on voit immédiatement apparaître les nombreuses limitations auxquelles elle se heurte :

- le programme ne peut pas toujours prendre en compte toutes les valeurs que peut prendre un critère (c'est le cas par exemple des grandes nomenclatures géographiques, à cause de leur longueur)
- on ne dispose pas toujours de la place mémoire nécessaire pour appliquer cette méthodes brutalement,
- le nombre d'opérations de calcul est grand par rapport à celui qu'on obtiendrait en raisonnant un peu pour améliorer ce procédé,
- enfin, les calculs de marge, de pourcentage, les calculs entre sous-cases, la présentation du tableau ne sont pas pris en compte par ce schéma simpliste et il faut savoir où les faire intervenir.

Nous allons donc essayer de décomposer en étapes de traitement un processus de tabulation qui prenne en compte les remarques que nous venons de faire.

3.1.) Ventilation brute fermée et génération de clés de tri (première étape).

Il est utile de distinguer les critères dont la liste de modalités, assez réduite, peut donc être connue à l'avance du programme, et les autres, dont on rencontrera les modalités au cours du déroulement du fichier, sans que leur liste soit connue du programme. On appelera les premiers critères fermés, les seconds critères ouverts, pour la suite de notre propos.

Pour les critères, on peut construire en mémoire un ensemble d'emplacements tels qu'on les a définis ci-dessus ; on effectuera les cumuls à l'intérieur de ces emplacements, au fur et à mesure du déroulement du fichier d'enquêtes tant que la valeur des critères ouverts ne change pas. Chaque fois qu'on rencontrera une valeur différente pour au moins un des critères ouverts - on parlera alors de rupture - on écrira sur un fichier intermédiaire un enregistrement contenant tous ces emplacements de mémoire, ainsi que lesvaleurs correspondantes des critères ouverts, qui formeront des clés de tri pour l'étape suivante.

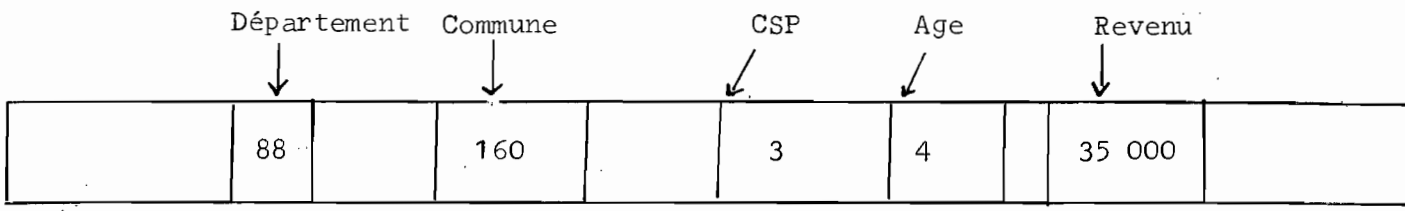
On voit que cette méthode est théoriquement applicable quelle que soit la manière dont est ordonné le fichier d'enquête : cependant, il apparaît qu'il devra être ordonné de la manière la plus cohérente possible avec les critères ouverts, et, si c'est possible, on triera auparavant le fichier d'enquête sur ces critères. En effet, si le fichier se présente en désordre, il y aura dans le fichier intermédiaire de nombreux enregistrements, chacun contenant un grand nombre d'emplacements à blanc. Mais la place nécessaire en mémoire est toujours la même ; elle est connue à l'avance compte tenu du nombre de modalités des critères fermés et de la longueur des critères ouverts.

Au cours de la même étape, on peut aussi préparer les opérations de récapitulation qu'on peut vouloir appliquer sur les critères ouverts : ainsi, pour obtenir un total sur un critère ouvert, il suffira de générer deux enregistrements identiques pour chaque valeur des critères, l'une correspondant à la modalité du critère, l'autre à une modalités "total" : l'étape suivante, qui est un tri-cumul, provoquera le cumul de toutes les ventilations correspondant à la modalité "total". Cette méthode est donnée ici à titre indicatif : il y a lieu de l'employer avec prudence, car elle peut dégrader les temps de traitement en multipliant les temps d'écriture et de lecture sur les fichiers intermédiaires.

Exemple :

L'exemple simpliste que nous donnons permet d'illustrer les définitions introduites dans ce paragraphe et montre comment s'enchaînent les étapes que nous préconisons pour la tabulation.

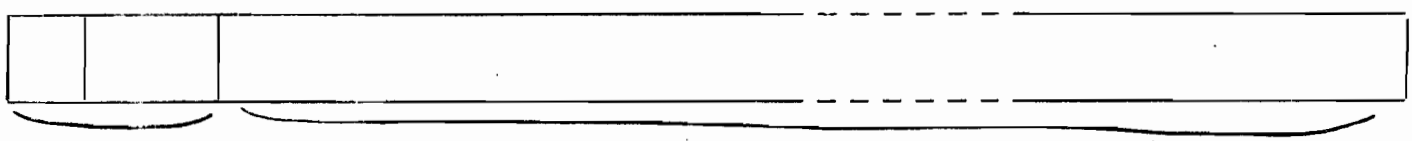
Fichier d'enquête :



- département, commune : critères ouverts
- CSP : critère fermé à 8 modalités
- âge : critère fermé à 10 modalités (tranches)

Supposons qu'on veuille réaliser un tableau croisant CSP et âge pour chaque commune, chaque département, et pour le fichier entier ; on veut dans chacun des tableaux le nombre d'individus, le revenu, et le revenu moyen ; on veut ces données pour chaque croisement des critères CSP et Age, pour chaque CSP tous âges réunis, pour chaque âge toutes CSP réunies, et pour l'ensemble de la population.

Les enregistrements du fichier intermédiaire seront de la forme :



clé de tri
dépt. commune

8 x 10 x 2 = 160 nombres

correspondants au croisement des deux critères fermés, avec deux emplacements par couple de valeurs des critères, l'un pour le comptage des individus, l'autre pour le cumul de leurs revenus.

Dans le cas où l'enregistrement dessiné ci-dessus était seul en rupture, il donnerait lieu à la génération de trois enregistrements pour le fichier intermédiaire :

a)

88	160	0	0	0		1	35 000		0	0

pour le tableau relatif à la commune 88 160
les emplacements non nuls correspondent à la trosième modalité du CSP et à la quatrième modalité d'âge

b)

88	...	0	0	0		1	35 000		0	0

pour la récapitulation relative au département 88

c)

..	...	0	0	0		1	35 000		0	0

pour la récapitulation relative au fichier entier

3.2.) Tri-Cumul (deuxième étape)

Si l'on effectue un tri-cumul sur le fichier intermédiaire obtenu, le résultat sera un fichier où il n'y aura plus qu'un enregistrement par valeur possible de la ventilation ouverte, c'est-à-dire pour chaque combinaison possible des critères ouverts, y compris les combinaisons qui correspondent aux récapitulations.

Dans notre exemple, on trouvera ainsi un seul enregistrement par commune, par exemple :

88	160	421	9 000 000		1253	35 000 000	352	7000 000

On aura aussi un enregistrement pour chaque département, et un enregistrement pour la récapitulation générale.

Remarque :

Cette méthode s'applique particulièrement, bien que cela n'apparaisse pas sur l'exemple, au cas où on réalise tout un ensemble de tableaux différents sur le fichier d'enquête : la clé de tri comportera alors une partie qui identifie le tableau concerné par l'enregistrement.

3.3.) Ventilations complètes-fermées, calculs entre sous-cases :

3ème étape.

Au cours de l'étape suivante, on peut travailler en mémoire pour "compléter" la ventilation fermée, c'est-à-dire, dans notre exemple. calculer les marges du tableau : on dispose en effet, en mémoire, en relisant le fichier issu de l'étape précédente, de tous les éléments nécessaires pour :

- calculer le nombre d'unités et cumuler les revenus de chaque CSP tous âges confondus,
- calculer le nombre d'unités et cumuler les revenus de chaque tranche d'âge toutes CSP confondues,
- calculer le nombre d'unités et cumuler les revenus de tous les individus d'une ventilation ouverte donnée.

C'est également au cours de cette étape que l'on effectuera les calculs entre sous-cases : dans notre exemple, il suffira de diviser le revenu par le nombre d'individus pour obtenir le revenu moyen.

Il faut selon les cas faire ce genre de calcul avant ou après le traitement des marges : dans le cas d'une moyenne, ce calcul s'effectue évidemment après l'obtention des marges.

Cette étape donne lieu à l'écriture d'un nouveau fichier intermédiaire, dont le dessin est cette fois modifié : pour chaque ventilation ouverte, nous avons en effet dans notre exemple :

$$(8 + 1) \times (10 + 1) \times 3 = 297 \text{ nombres.}$$

3.4.) Dernière étape : édition.

Arrivé à ce stade du traitement, toutes les informations figurant dans les tableaux ont été calculées. Il est bon de consacrer un programme particulier à l'édition, c'est-à-dire à l'écriture sur le papier continu utilisé en informatique des tableaux demandés.

Se posent en effet un certain nombre de problèmes qu'il est bon de traiter séparément des calculs, car leur complexité, d'un autre ordre, n'en n'est pas moindre :

- découpage du tableau en fonction des dimensions respectives de celui-ci et des feuilles de papier,
- adjonction de titres, des modalités ou des intitulés des critères,
- écriture des nombres dans un format approprié,
- détermination de la mise en page (critères présentés en ligne, en colonne, etc...).

Séparer les fonctions d'édition des fonctions de calcul permettra donc une mise au point séparée des programmes, une modification de la présentation des tableaux sans avoir à refaire la tabulation ; c'est donc une précaution qu'il est souhaitable de prendre.

Il faut aussi souligner que les fonctions d'édition sophistiquées des tableaux ne sont pas toujours nécessaires, et qu'elles posent de délicats problèmes informatiques : c'est pourquoi, pour les petits dépouillements rapides, il sera souvent plus efficace d'effectuer une impression sommaires des résultats et de prévoir des liseuses, bandes de papier par exemple à coller à côté des chiffres imprimés par le programme.

IV - EVALUATION D'UN LOGICIEL DE TABULATION

Nous nous contenterons dans ce paragraphe de fournir une grille qu'il peut être intéressant de remplir si on doit évaluer un logiciel de tabulation. Cette évaluation devant permettre de voir si ce logiciel permet de réaliser les tableaux désirés compte tenu du fichier de données ; et aussi si on peut l'implanter dans l'ordinateur dont on dispose.

On peut distinguer :

- Les paramètres d'exploitation

  - Type de machine sur lequel le système est opérationnel et mode de supervision.
  - Taille de la mémoire centrale nécessaire.
  - Mode d'utilisation (batch, conversationnel)
  - Langages de programmation (évolués) utilisés pour écrire le logiciel (compilateurs nécessaires), nombre d'instructions.
  - Assembleurs utilisés
  - Unités périphériques nécessaires :
    - Lecteur
    - Imprimante
    - Disques
    - Bandes
    - Autres
  - Processeurs nécessaires à l'utilisation.
  - Performances (souvent difficile à évaluer, car varie d'une installation à une autre)
  - Type et structure du fichier de données (physique)
  - Nature du système (programme général, ou compilateur
  - Accès au système
    - langage à écriture libre
    - langage à écriture zonée
    - appel de fonction par mots-clés
    - cartes paramètres
  - Possibilités de reprise en cas d'incident d'exploitation.

- Possibilités du système et limitations

. Structure logique du fichier de données (plat ou hiérarchisé)

. Nombre de type niveaux hiérarchiques

. Nombre de critères croisés possibles

. Type des quantités cumulées (entiers ou flottants)

. Valeur maximale des quantités cumulées

. Possibilité de filtres ou de critères d'appartenance

. Possibilité d'opération entre tableaux

. Possibilité d'opération entre cases d'un tableau

. Possibilités diverses (récapitulations, pourcentages, tranche, etc...

. Taille maximum d'un tableau

. Présentation des tableaux

édition standardisée
modification possible de l'édition
libellés

. Possibilité de reprise de fichiers de tableaux

. Possibilité de calcul d'analyse statistique sur un tableau

- Conditions d'utilisation du logiciel

Les critères qui suivent sont plus subjectifs mais demandent a être examinés avec soin :

. . Difficulté d'apprentissage du langage de tabulation ou d'écriture des paramètres

. Le fournisseur organise-t-il des stages de formation

. Qualité de la brochure et du matériel pédagogique fournis

. Conditions commerciales d'acquisition et d'utilisation du produit, de la maintenance

Enfin on peut signaler que l'implantation d'un système un peu complexe est toujours délicate et nécessite généralement la présence d'un spécialiste du produit.

On passera donc en revue tous ces points selon les situations particulières, on pondérera l'importance des différents critères de choix en se souvenant que certains sont redhibitoires.

5 - GESTION DES TABLEAUX

Le troisième paragraphe présentait un schéma de construction pour les programmes de tabulation : le but de ces programmes était de produire des tableaux, a priori imprimés sur papier. En fait, il faut bien prendre conscience que ces tableaux sur papier ne peuvent plus qu'être étudiés, publiés... alors que, sur support informatique, ils pourraient servir à de nombreux traitements.

C'est pourquoi il faut prendre garde, sous peine d'avoir à en effectuer de nouveau la saisie -avec les risques d'erreurs inhérents à cette opération- ou de repartir du fichier enquête -avec le coût supplémentaire induit par cette opération- de définir, en plus des sorties imprimées, des fichiers de tableaux.

Définir ces fichiers est plus délicat que définir le dessin des fichiers d'enquête, car tous les questionnaires ont la même allure, alors que les tableaux sont variés : mais il est utile cependant de le faire, pour les raisons que nous venons d'évoquer ci-dessus. Les logiciels généraux prévoient en général une façon de gérer les tableaux sur fichiers ; en cas de programmation spécifique, on pourra stocker les tableaux sous forme de flots de cases repérés par un numéro de tableau et la valeur des codes ouverts, de façon à se ramener à un dessin d'enregistrement standard pour ce fichier de tableaux. On peut aussi, en prenant des précautions, travailler directement sur un fichier contenant les tableaux prêts à être imprimés (un enregistrement contenant une ligne d'impression), mais cette façon de procéder suppose que l'on dispose d'un autre fichier contenant des descriptifs de tableaux, et est beaucoup plus complexe à mettre en oeuvre.

Quelle que soit la solution retenue, donnons deux exemples de traitements que l'on peut effectuer à partir de tableaux déjà calculés.

5.1 - Récapitulations de tableaux sur zones géographiques.

Si l'on a à produire des tableaux relatifs à des échelons géographiques d'importance variée, on aura intérêt à ne produire par l'opération de tabulation proprement dite que les tableaux relatifs à l'échelon le plus fin : il sera plus économique, pour produire les tableaux relatifs aux échelons supérieurs, de cumuler les résultats des échelons les plus fins.

Ceci est valable pour les échelons "emboîtés" : mais on peut généraliser, et, à condition de disposer des tableaux pour un échelon "de base" qui constitue un dénominateur commun des zonages administratifs divers -en France, la commune- on pourra produire à la demande, sans retour au fichier d'enquête, et sans même disposer dans le fichier d'enquête de tous les codes géographiques possibles, des tableaux pour des zones quelconques. Il suffira pour cela de constituer un répertoire des zonages, consulté par le programme de récapitulation.

5.2 - Production de tableaux par calculs entre tableaux.

Certains tableaux peuvent être obtenus directement par calculs entre plusieurs autres tableaux, ou se déduire d'un autre tableau, sans retour au fichier d'enquête. En étudiant cette possibilité dès que l'on dispose du plan de tabulation, ou au moins en se la réservant par définition d'un "fichier de tableaux" dans le cas où les tableaux à produire ne sont pas encore tous définis, on réalisera donc une importante économie de traitement. De plus, il existe des cas où certains tableaux calculés à partir de deux autres tableaux n'auraient pas pu être produits directement à partir du fichier d'enquête.

Par exemple, si l'on dispose d'un tableau cumulant les dépenses des ménages suivant certains critères et d'un autre cumulant les biens consommés suivant les mêmes critères, on obtiendra un tableau contenant le prix auquel reviennent ces biens suivant les mêmes critères par division des deux tableaux case à case.

Des exemples moins simples peuvent faire intervenir des tableaux n'ayant pas exactement les mêmes critères : il suffit que les critères de tabulation correspondant au tableau résultat soient un sous-ensemble des critères des tableaux qui interviennent dans le calcul.

Il existe des logiciels généraux dont la fonction est de réaliser des calculs entre des tableaux figurant sur des fichiers d'un standard défini : de tels logiciels complémentaires des logiciels de tabulation et se plaçant "en aval" de ceux-ci, peuvent présenter un intérêt, bien que leur usage ne soit pas aussi répandu que celui des logiciels de tabulation.

Chapitre 10

L'information en sortie

Rédacteur : A. BREAS

Etat de la rédaction : Plan détaillé

Plan du Chapitre :

- Tableaux brouillon,
- Tableaux de diffusion,
- Tableaux destinés à la publication,
- Courbes, graphiques et cartographie,
- Photocomposition.

CHAPITRE 10

L'INFORMATION EN SORTIE

On distinguera les divers types de "sorties" :

- tableaux "brouillon",
- tableaux de diffusion,
- tableaux "publication",
- courbes, graphiques, cartographie,
- photocomposition.

A - TABLEAUX "BROUILLON"

Destinés au seul statisticien-demandeur (tableaux de contrôles, tableaux d'étude destinés à affiner les demandes suivantes etc...)

Impératif :

Production la plus rapide possible (la rapidité prime sur la "qualité", il suffit que ces tableaux puissent être interprétés par le statisticien-demandeur).

B - TABLEAUX "DIFFUSION"

Seront consultés par un large public, y compris de non spécialistes.

Impératifs :

a - Faciliter la compréhension

- utilisation de préimprimés,
- utilisation d'intitulés en clair explicites,
- fabrication de "liseuses" pour accompagner les tableaux comportant des intitulés en code ou très abrégés.

b - Faciliter la consultation

- en référençant chaque page dans une liasse,
- en référençant les diverses pages d'un même tableau,
- en créant des index (table des matières : identification du tableau - n° des pages) automatiquement au cours de l'édition (ou à défaut à la main),
- (ordre des tableaux ?).

c - Faciliter la production

- sorties d'édition sur bande "spool" + programme d'impression avec possibilité de "reprises" (pour nouvelle sortie des pages endommagées ou peu lisibles pour pouvoir fractionner les sorties d'un très gros volume etc...).

.../...

d - Faciliter le "dispatching"

- en prévoyant un découpage par destinataire (au niveau du tri des tableaux, de leur identification, de l'édition),
- en utilisant la table des matières,
- en adaptant le support de sortie en fonction du nombre d'exemplaires à produire :

  . sortie sur papier simple ou multiple (problème de lisibilité au-delà de 3 exemplaires selon le papier utilisé),

  . sortie sur plaques offset (tirage par l'imprimeur ---> 100 exemplaires),

  . sortie sur papier puis microfilmage en continu par caméra dynamique (reproduction du microfilm très rapide selon le matériel de "contretype" utilisé),

  . sortie directe sur film ou fiches (systèmes COM - possibilité de superposer un fond de page préimprimé aux tableaux, ou selon le COM de "tracer des fonds de page programmés").

e - Assurer une excellente qualité d'impression

- utilisation de "rubans" d'imprimante particuliers,
- étude du grammage et de la qualité du papier en fonction des performances de l'imprimante utilisée,
- précautions à prendre au niveau de l'imprimante :
  . réglage, nettoyage, test d'usure des caractères, etc...
- précautions à prendre au niveau du traitement du papier.

f - Pouvoir retrouver ce qui a été fait et archivé

- prévoir un système d'identification des tableaux :
  . numéros de source,
  . numérotation des tableaux à l'intérieur d'une source,
  . distinguer par une numérotation "articulée" les différentes "familles de tableaux", etc...
- produire des tableaux synoptiques décrivant les tableaux produits (champ, niveaux géographiques, critères ligne-colonne-totalisation) et associant à chaque description les références (n° source -n° tableau)
- enrichir les fichiers "tables de matières" des références de stockage des tableaux (n° de bobine de film, n° de microfiche, n° armoire, etc...)

.../...

- prévoir de gérer ces différents fichiers (tri selon les différants critères d'accès :
  - niveaux géographiques,
  - ou numéro de tableau,
  - ou numéro de famille, de source, etc...
    production d'index édités)

- prévoir une banque d'index des tableaux produits (ensuite des travaux).

C - TABLEAUX PUBLICATION

Remis à l'imrimeur : nécessité de contacts préalables avec l'imprimeur pour établir sans ambiguité les spécifications des documents qui lui seront remis et du type de traitement qu'il leur fera subir :

a - photographie de listings

- des tableaux imprimés tels quels (ils doivent être parfaitement compréhensibles - on devra éventuellement tirer des traits : voir pb. de l'encrage),

- de tableaux enrichis de titres et intitulés composés sur machine à écrire ou varituper, collés sur les listings (pb. papier et collage),

b - photographie de listings avec ajout de têtières par l'imprimeur

- convient par exemple pour les tableaux du type "inventaire",

- pb du niveau de la mise au point préalable des têtières pour ne pas avoir de problème de cadrage,

c - photographie de listing avec superposition d'un fond de page par l'imprimeu

(Le fond de page est dessiné sur un translucide superposé au tableau imprimé au moment de la photographie).

Etude rigoureuse du contenu du fond de page, du cadrage des données, etc... (méthode à suivre : dessin sur pages de petits points, étude des problèmes de bandes pilotes, etc...).

d - composition traditionnelle de tableaux à partir d'un "manuscrit" constitué manuellement (ou avec insertion de fractions de tableaux découpées et collées) par le statisticien.

D - COURBES, GRAPHIQUES, CARTOGRAPHIE

Nécessitent : Logiciels se greffant si possible en aval du logiciel de tabulation + matériel : traceurs électro-mécaniques
ou système COM (traceurs/unités mixtes, traceurs imprimantes sur film ou fiches).

E - PHOTOCOMPOSITION

- différents modes d'utilisation avec avantages et contreparties éventuelles

- champ d'utilisation

Chapitre 1 1

La planification des travaux et la documentation des traitements : le cahier des charges

Rédacteur : J.P. LACHIZE

Etat de la rédaction : définitive

Plan du Chapitre :

I) Introduction,

II) La place du cahier des charges dans le dépouillement,

III) La constitution du cahier des charges, son caractère évolutif,

IV) Le contenu du cahier des charges

Avertissement -

La fin du chapitre 11 est actuellement manquante. Les alinéas ou paragraphes manquants sont les suivants :

IV.1.f. - Les prévisions de charge

- charges informatiques,
- charges manuelles,

IV. 2 - Les dossiers de phases

V. Le cahier des charges : documentation des traitements

I - INTRODUCTION

Un cahier des charges, est-ce utile ? Des statisticiens, des informaticiens, s'ils ont déjà ou non une expérience et donc une opinion sur ce document, peuvent donner des réponses des plus variées. Ces mots "cahier-des-charges" recouvrent souvent pour chacun d'eux une réalité différente, parfois pénible, car ressentie comme inutile : un papier de plus à écrire qui devient vite obsolète. Est-ce donc seulement une procédure administrative ou autre chose ?

L'idée principale de ce chapitre est de faire du cahier des charges un réel instrument de travail, dont l'existence détermine la bonne conduite des travaux de dépouillement d'enquête. Ce cahier des charges sera aussi la base d'une documentation complète, cohérente et facile à consulter.

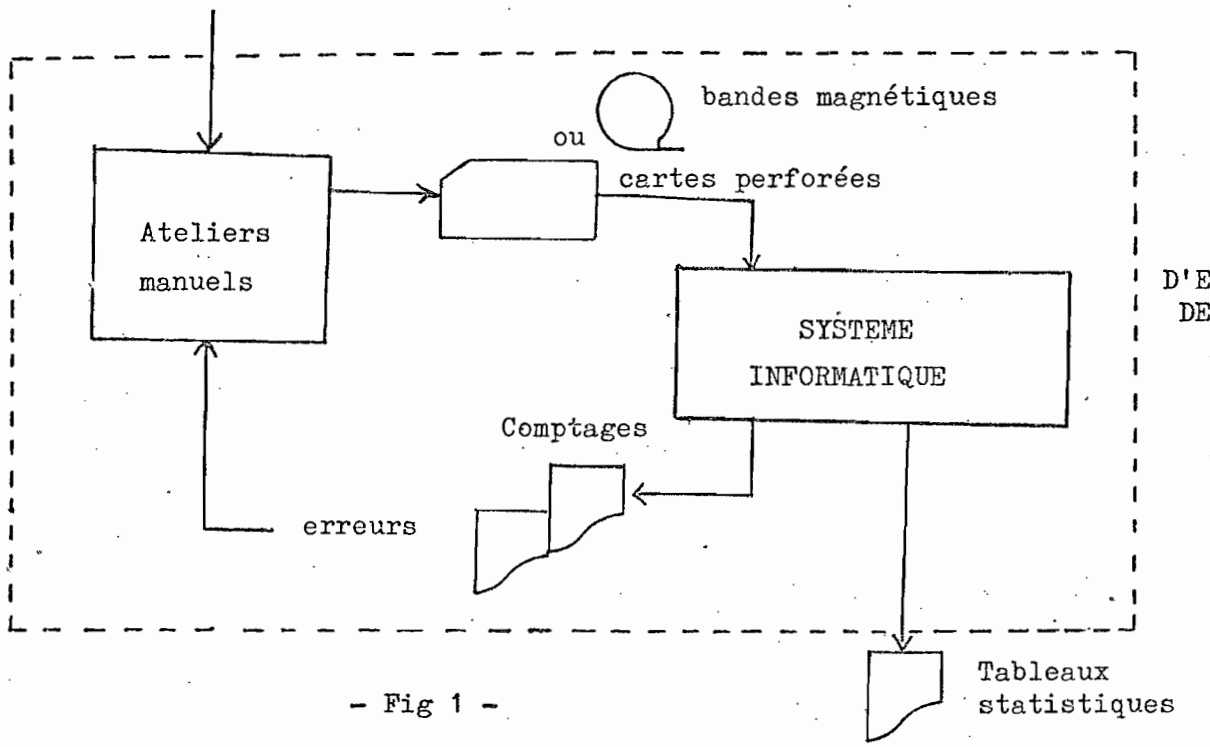
Que ce document s'appelle cahier des charges ou qu'il s'appelle autrement, ce n'est pas là l'essentiel. L'essentiel est qu'il existe bien à sa place dans le contexte d'un dépouillement d'enquête et que son contenu soit parfaitement défini.

Il ne s'agira pas cependant dans ce chapitre d'imposer des règles, d'imposer une forme de cahier des charges... mais beaucoup plus d'inciter à la réflexion sur les relations entre les différentes personnes qui travaillent sur un dépouillement et sur les procédures de gestion à employer.

II - LA PLACE DU CAHIER DES CHARGES DANS LE DEPOUILLEMENT D'UNE ENQUETE

Le cahier des charges est d'abord l'ACCORD écrit entre les différents partenaires (statisticiens, informaticiens...) aussi bien sur les objectifs à atteindre et les moyens nécessaires que sur la nature exacte des traitements à effectuer.

Il peut s'agir de traitements informatiques : programmes de contrôle, de codification, de tabulation... Il peut s'agir de traitements manuels : chiffrement, saisie...



- Fig 1 -

Ce schéma décrit le système d'exploitation de l'enquête une fois qu'il est réalisé. Le système informatique ne représente que l'exploitation sur ordinateur des programmes. Ce système est piloté par des gestionnaires.

Cette approche globale et toute la formalisation qui va suivre, conviennent plus à une enquête lourde, genre recensement, qu'à une petite enquête réalisée parfois entièrement par le statisticien lui-même, dans ce dernier cas, il faut donc adapter ce qui suit.

Les différents partenaires

On considère qu'il y a une nette séparation entre l'aspect purement statistique et l'aspect gestion - production et que cette séparation se retrouve au niveau des personnes. Cette approche toute théorique ou conforme à la réalité suivant les cas, permet de mieux cerner le problème.

D'un côté le statisticien agissant en qualité de demandeur (de client) et de l'autre un responsable de la production et tous ceux qu'il coordonne :

- les informaticiens : analystes, programmeurs,
- les responsables de la mise en place des ateliers manuels,
- les gestionnaires.

Dans les faits, le statisticien exerce souvent la fonction de responsable de la production, ou du moins en partie, les informaticiens sont alors encadrés par un chef de projet qui n'élabore que le système informatique.

Que ce responsable de la production soit un informaticien est une autre solution, il ne limitera pas ainsi sa compétence au domaine informatique mais l'étendra à l'ensemble du système d'exploitation de l'enquête : les traitements informatiques ont souvent de fortes répercussions (messages d'anomalies) sur les tâches manuelles et inversement (chiffrement). Il faut donc que le responsable de la production ait une bonne connaissance de la logique de l'ordinateur et de ses possibilités.

Dans tous les cas, le cahier des charges doit permettre de clarifier les responsabilités de tous ceux qui travaillent sur le dépouillement de l'enquête.

Les objectifs

Les objectifs s'expriment en délais de réalisation et en qualité des traitements à effectuer. Par exemple, si l'on se propose de dépouiller une enquête en 6 mois, la qualité des traitements (contrôles, redressements...) doit être fixée en conséquence.

Il faut cependant se méfier de certains sacrifices de qualité faits dans la précipitation pour réduire les délais et qui ont souvent des effets inverses.

Le cahier des charges contiendra donc une description précise des objectifs à atteindre, en particulier sous forme de planning pour les délais.

Les moyens

Le cahier des charges contiendra une estimation des moyens nécessaires au dépouillement de l'enquête. Ces moyens sont, bien sûr, fonction de la nature des traitements. Une estimation valable ne peut être faite que si ces traitements sont suffisamment bien définis au moment de la constitution du cahier des charges.

Délais et moyens seront prévus une première fois en début de travail, puis ensuite ils seront régulièrement suivis au cours d'une réunion mensuelle des différents responsables afin d'éviter les mauvaises surprises.

La nature des traitements

Une première approche, quelque peu grossière, permet de considérer que trois étapes se succèdent dans le temps :

C - la conception : * étude du cheminement de l'information, du questionnaire à la sortie des tableaux, détermination des traitements informatiques et manuels : contenu statistique (contrôles, redressements, codifications...), principales chaînes de programmes, structure et contenu des fichiers... etc.

* estimation des moyens nécessaires : informatiques et manuels.

* planification de la réalisation.

RM - la réalisation : des chaînes de programmes (analyse (1), écriture, tests), leur documentation (dossiers de programme et d'exploitation et la mise en place des ateliers manuels (organisation, formation du personnel, production des instructions et documents nécessaires).

E - l'exploitation : traitements informatiques et manuels : des données du premier questionnaire au dernier tableau.

(1) l'analyse faite à ce niveau, est purement technique et ne concerne que les informaticiens.

Le cahier des charges est le document de synthèse des études de conception, il se constitue au fur et à mesure que ces études progressent.

Il contient en conséquence toute l'information nécessaire à la réalisation et échangée entre le (ou les) statisticien(s) et la production.

R E M A R Q U E S :

1 - L'importance des travaux fait qu'en général, il y a chevauchement de ces trois étapes : on peut en être à la conception des tableaux et les questionnaires être en cours de saisie (exploitation).

2 - La conception n'est jamais parfaite, il y a toujours au niveau de la réalisation des remises en cause de détails.

III - LA CONSTITUTION DU CAHIER DES CHARGES : SON CARACTERE EVOLUTIF

Le statisticien a l'initiative des travaux. Il est absolument nécessaire qu'au départ, il constitue un dossier statistique suffisamment complet pour servir de base aux études de conception :

- objectifs de l'enquête, résultats recherchés,
- méthode de sondage (lorsqu'il y en a un),
- méthode de collecte (ponctuelle, répétée...),
- modèle de questionnaire, parties obligatoires et facultatives,
- variables retenues pour le chiffrement, modalités envisagées (numérique, alphabétique, alphanumérique possibilité d'un blanc),
- codification et contrôles envisagés,
- corrections envisagées : automatique, manuelle. Importance en volume de ces corrections,
- les calculs,
- les tabulations,
- ... etc.

Ce dossier statistique permettra à un groupe d'étude formé du responsable de production, d'informaticiens - analystes, de spécialistes des traitements manuels, et qui auront comme principal interlocuteur le statisticien, de concevoir les traitements.

Il est important que ce travail soit le fruit d'une équipe et non d'une individualité : d'abord par sécurité, puis parce que la conception en sera souvent meilleure et enfin, pour faciliter l'insertion des personnes qui réalisent le dépouillement, elles se sentiront en général plus motivées si elles ont participé à la conception.

.../...

Le cahier des charges sera rédigé par ce groupe d'étude, à l'exclusion du statisticien. Ce cahier des charges reprendra entièrement la note du statisticien, ce qui est nécessaire pour que ce dernier se rende compte s'il a été compris.

En effet, pour s'assurer qu'une information est bien passée, par exemple entre un statisticien et un informaticien, il ne suffit pas de faire une note, de faire une réunion ou de téléphoner. Ces façons de procéder sont certes indispensables, mais si elles sont nécessaires elles ne sont pas suffisantes. Le cahier des charges rédigé par le "récepteur" de cette information, permet à l'"émetteur" de contrôler que son information a été bien reçue, sans déformation.

Il est toujours délicat de prétendre que le récepteur capte mal, plutôt que l'émetteur émet mal ("il y a eu une note..."! réponse "elle n'était pas claire"!).

Il faut aussi éviter l'éparpillement de l'information, d'où l'intérêt de regrouper toute l'information échangée entre le statisticien et la production dans un document unique.

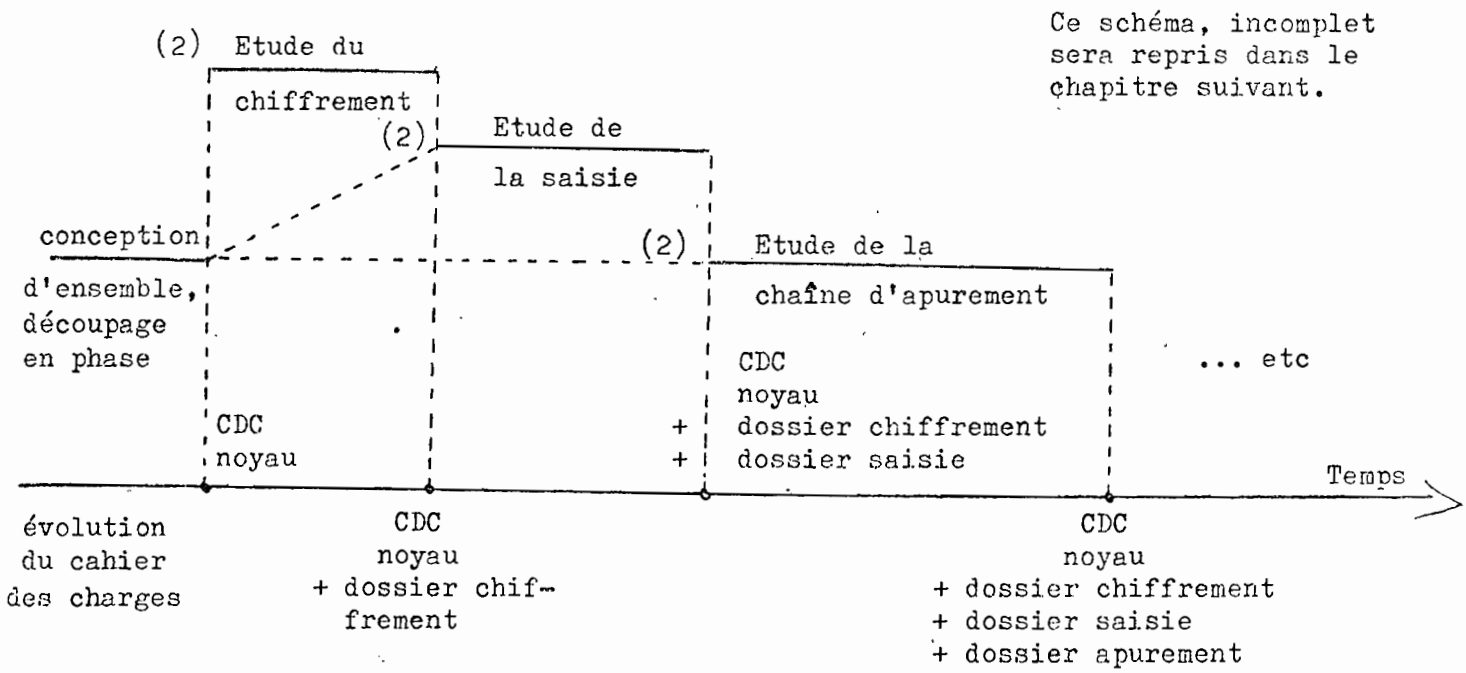
Le cahier des charges doit pouvoir être mis à jour (1) et il sera géré par le responsable de la production qui disposera de l'exemplaire de référence.

(1) Côté pratique : le document peut être conservé dans un classeur pour feuilles perforées, les pages numérotées de 10 en 10 et datées. Ce qui permet des insertions et de conserver les pages périmées comme historique. Une modification (exemple : suite à une demande écrite du statisticien) conduit à refaire les pages concernées : les pages décrivant le traitement, et si nécessaire, avec l'accord des différentes parties, les pages d'estimation des moyens et de plannings. Ces pages sont diffusées à toutes les personnes qui ont un exemplaire du CDC et en particulier au demandeur de la modification (contrôle).

Le cahier des charges sera établi progressivement, au fur et à mesure qu'avance la conception.

A la conception d'ensemble correspondra la partie principale du cahier des charges (noyau) où apparaîtra un découpage du projet suivant ses grandes fonctions (saisie, apurement, tabulation... etc), la conception de chacune de ses fonctions (ou phase) (1) correspondra à un dossier d'analyse, inclusdans le cahier des charges.

Exemple :



(1) La notion de phase sera définie dans le chapitre suivant

(2) L'étude proprement dite a commencé bien avant cette date, la conception d'ensemble (noyau) nécessite d'avoir déjà défini et évalué chacune de ces parties, il s'agit plus précisément du début d'une étude détaillée qui permet de fixer (par écrit) les traitements statistiques à faire, au niveau le plus fin (code). Le dossier, résultat de cette étude, contient tous les éléments nécessaires à la réalisation des programmes ou à la mise en place des ateliers.

Le cahier des charges ainsi défini a deux facettes, l'une tournée vers le demandeur (statisticien), l'autre vers la réalisation.

Il matérialise l'accord conclu entre différentes personnes sur les travaux à faire, il doit donc être écrit dans un langage clair pour tous, bannissant les termes trop typiquement informatiques ou statistiques.

Pour faciliter la compréhension, il sera souvent indispensable que le statisticien fasse quelques séances d'information sur son travail, particulièrement si le domaine des enquêtes statistiques est peu familier aux informaticiens.

IV - LE CONTENU DU CAHIER DES CHARGES

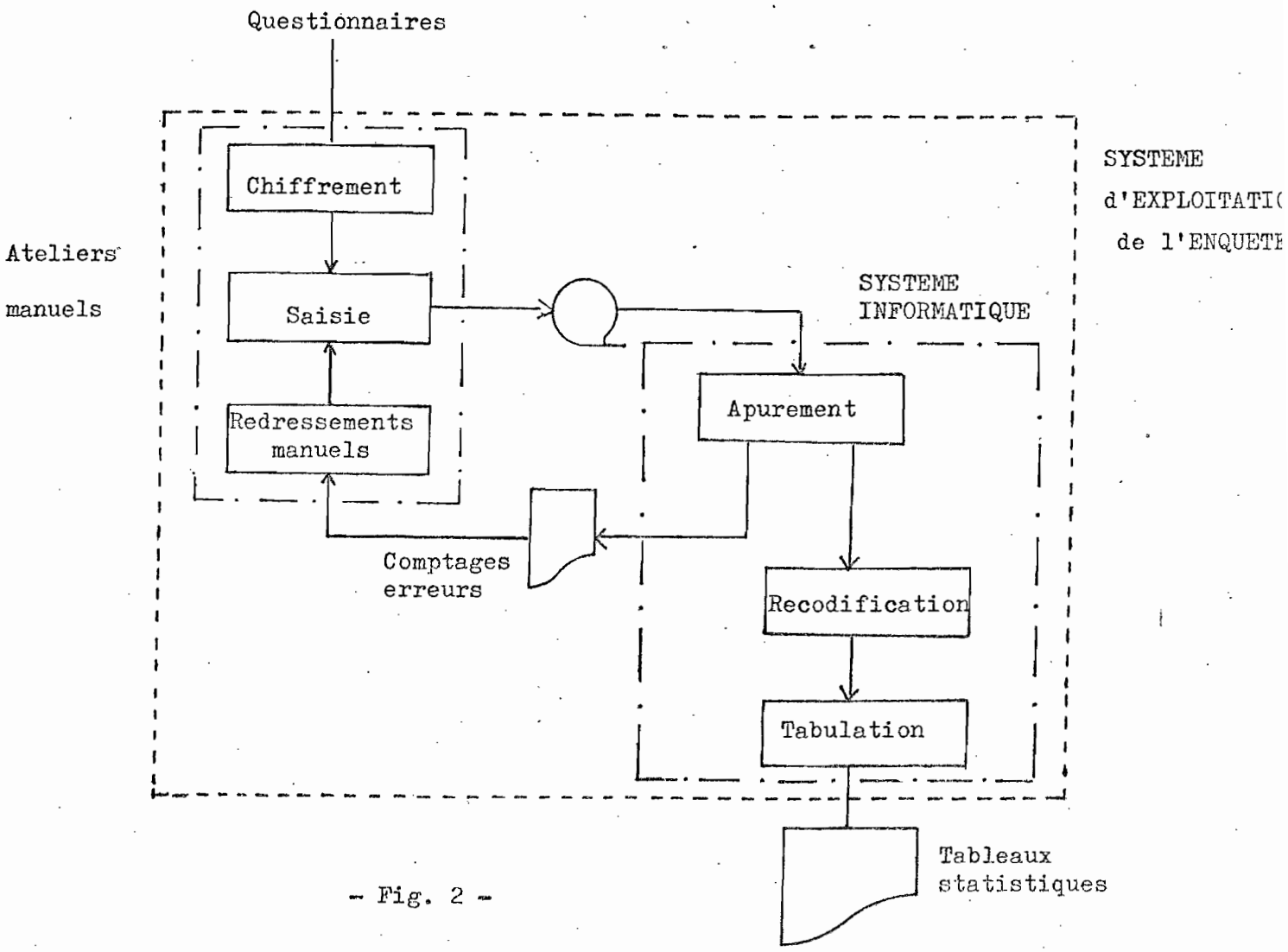
IV - 1 - Le noyau

Il contient :

a - une présentation générale de l'enquête avec les objectifs et résultats recherchés.

b - un schéma global du dépouillement de l'enquête en faisant apparaître les principales fonctions (chiffrement, saisie, apurement... etc).

Exemple :



- Fig. 2 -

c - un découpage du projet en phase

Tout le projet ne peut être conçu dans ses détails, les programmes réalisés, les ateliers mis en place, en même temps. Il est donc nécessaire de découper les travaux en ensembles homogènes et faisables en un laps de temps raisonnable (quelques mois).

Pour faciliter ce découpage, fait à partir de la décomposition fonctionnelle (paragraphe précédent), les travaux à effectuer peuvent être répartis par nature, en dégageant pour chaque fonction la partie conception (C), la partie réalisation ou mise en place d'atelier (RM) et la partie exploitation des données (E).

A la constitution du noyau du CDC, la REFLEXION SUR CHAQUE PHASE DOIT ETRE SUFFISAMMENT AVANCEE POUR NE PAS REMETTRE EN CAUSE plus tard l'ENSEMBLE DES MOYENS PREVUS (§ IV.1.F) ET LES OBJECTIFS DEFINIS COMME IMPERATIFS (§ IV.1.a et IV.1.d).

Il faut donc aller plus loin qu'un simple découpage et amorcer une réflexion suffisante pour juger de l'importance de la phase, quitte une fois cette importance cernée, à l'abandonner pour la reprendre plus tard au moment de l'élaboration du dossier de phase.

Exemple :

Découpage en phase d'un dépouillement d'enquête dont les principales fonctions sont celles données paragraphe b, fig. 2

1 - Chiffrement

2 - Saisie

3 - Apurement

4 - Redressements manuels

5 - Recodification

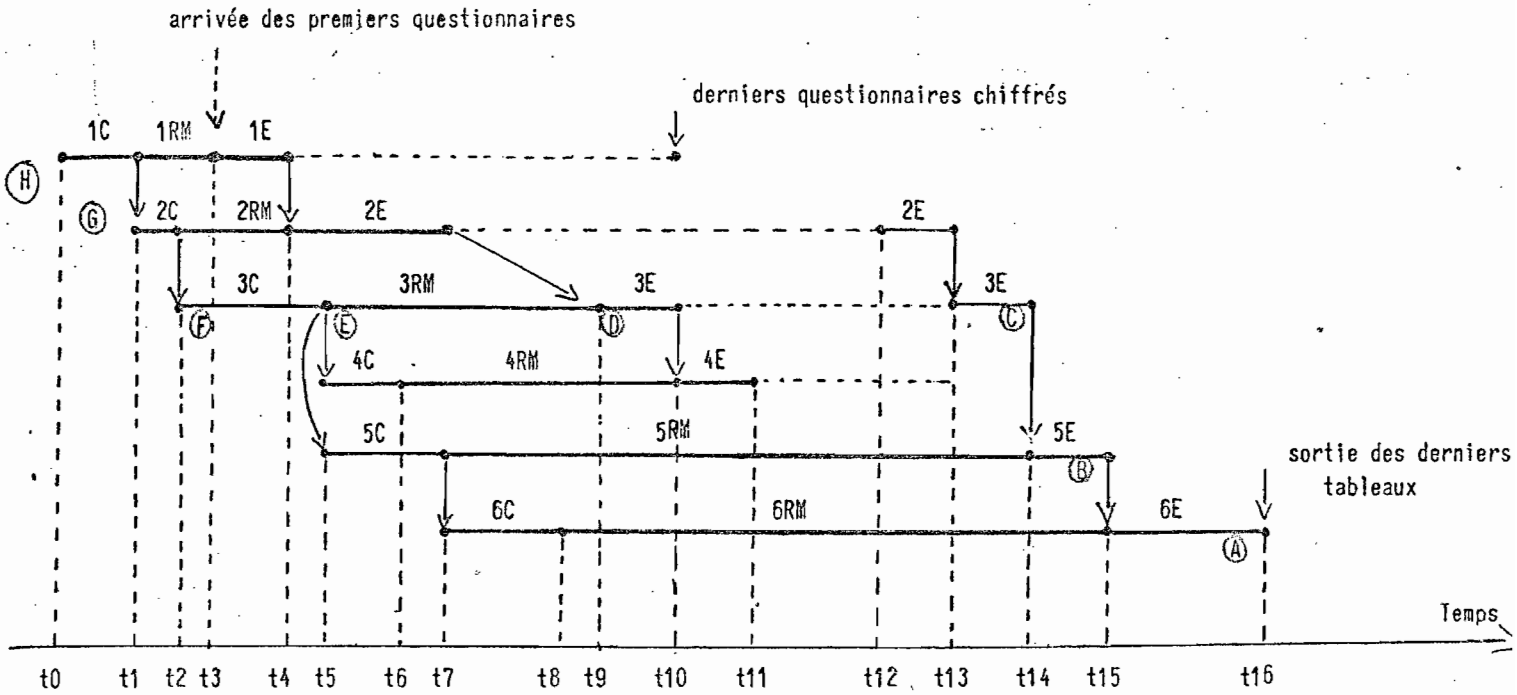
6 - Tabulation

Les phases :

	CONCEPTION	REALISATION	EXPLOITATION
1	PHASE 1C	PHASE 1RM	PHASE 1E
2	- 2C	- 2RM	- 2E
3	- 3C	- 3RM	- 3E
4	- 4C	- 4RM	- 4E
5	- 5C	- 5RM	- 5E
6	- 6C	- 6RM	- 6E

d - Le planning général : points de contrôles et contraintes

Exemple (suite) :



Les points de contrôle

t0 - le noyau du CDC, qui contient ce planning, est disponible

t1 - l'instruction de chiffrement est disponible
- analyse du programme de saisie

t2 - début programmation du programme de saisie
- début analyse chaîne d'apurement

t3 - arrivée des premiers questionnaires - chiffrement

t4 - saisie du premier lot de questionnaires chiffrés

t5 - le dossier de phase de la chaîne d'apurement est disponible
- étude de l'organisation des ateliers chargés des redressements manuels
- étude recodification

T6 - mise en place puis formation du personnel chargé des redressements manuels

t7 - le dossier de phase recodification est disponible
- étude de la tabulation

t8 - le dossier de phase tabulation est disponible

t9 - les premiers questionnaires ont été saisis, première exploitation de la chaîne d'apurement

t10 - les ateliers manuels reçoivent les premiers messages d'erreurs détectées par la chaîne d'apurement

t11 - renvoi à la saisie des premières corrections

t12 - saisie du dernier lot de corrections

t13 - dernier apurement du fichier

t14 - recodification du fichier

t15 - le fichier est propre, recodifié, prêt pour la sortie des tableaux

t16 - sortie des derniers tableaux

Les contraintes

- Les flèches indiquent sur le planning les opérations liées entre elles : la première doit être terminée pour que commence vraiment la seconde.

Exemple : il n'est pas possible d'étudier sérieusement le problème de la saisie tant que l'on ne sait pas ce qui sera chiffré sur le questionnaire.

- Le chemin critique.

Sur le planning, la distance entre deux points de contrôles n'est pas proportionnelle à l'importance du travail à effectuer, ce n'est qu'un délai.

.../...

Si l'on part du but à atteindre : la sortie des tableaux à la date t16 et que l'on remonte le temps en choisissant toujours la phase la plus importante quant au rapport travail à faire / délai, on détermine le chemin critique des dates qui risquent d'être les plus difficiles à tenir.

Dans l'exemple, le chemin critique est matérialisé sur le planning par les points A, B, C, D, E, F, G et H.

Corollaire : Tous les points de contrôle n'ont pas la même importance, plus on est prêt du chemin critique plus il faut veiller au respect des points de contrôles. Cette notion échappe souvent à des responsables qui suivent l'évolution d'un projet sans être directement impliqué dans son élaboration. Ils ont alors tendance à donner la même importance à chaque point de contrôle d'où des tensions et des erreurs.

Ce planning fait à la date t0 perd vite de sa valeur au fur et à mesure qu'avancent les travaux. Bien des dates ne sont pas tenues et le mieux à espérer (ce qui est suffisant) est le respect du chemin critique.

Son principal intérêt est de servir de référence aux plannings du court terme.

e - Les plannings du court terme

Cette partie du cahier des charges évolue avec l'avancement des travaux.

Le planning général recouvre une durée qui, dans bien des cas, est de l'ordre de l'année sinon plus. Il a de plus en plus de chance de se révéler inexact que l'on s'éloigne de la date t0, instant où il a été établi. Plus encore les dates intermédiaires (t6, t7, t8...) que la date finale (t16) qui, il faut l'espérer, a été choisie avec une marge de sécurité suffisante. Par contre, les dates (t1, t2, t3...) proches de la date de leur ~~évolution~~ évaluation ont plus de chance d'être tenues.

La gestion à court terme aura pour objet de fixer l'attention des responsables particulièrement sur les délais dont l'échéance est de 1 à 4 mois. Sur une période aussi courte, il doit être relativement facile de bien prévoir. D'où l'importance d'une réunion mensuelle pour repréciser dans un "planning à court terme" les points de contrôle du trimestre qui suit, et de voir au plus tôt les difficultés à venir.

Un planning à court terme peut être un tableau qui contient :

- toutes les dates de début et de fin de phase,
- pour le court terme, des dates plus précises, non plus au niveau de la phase, mais de chaque dossier, programme, exploitation... à réaliser.

Chaque réunion mensuelle donne lieu à la mise à jour du planning précédent.

PHASE	EXEMPLE DE PLANNING POUR LE SUIVI DES TRAVAUX	CDC établi le : 15/09/76	Révision faite le : 15/01/77	Révision faite le : 15/02/77		Date de réalisation
2C	Dossier de phase - saisie	01/02/77	15/02/77			13/02/77
2RM	Fin de la phase	15/05/77	-	01/06/77		
	programme de saisie directe sur bande	-	-	01/06/77		
	instruction de saisie pour les opératrices	-	-	15/05/77		
3C	Dossier de phase - Apurement	15/02/77	-			14/02/77
3RM	Fin de la phase	01/06/77	-	-		
	1er programme de contrôle	-	-	01/05/77		
	2ème programme de contrôle	-	-	15/05/77		
	programme d'édition	-	-	15/06/77		
	dictionnaire de messages d'anomalies	-	-	01/06/77		
2E	Saisie des premiers questionnaires	01/06/77	-	-		
⋮	⋮	⋮	⋮	⋮		⋮
etc	etc	etc	etc	etc		etc

<u>R E M A R Q U E</u> : C'est une fois le dossier de phase constitué que l'on connaît et que l'on peut fixer des dates pour les différents travaux qu'il recouvre.

A N N E X E 1

Présentation du logiciel général S P S S

Avertissement : Ce texte est une traduction d'un article paru en anglais dans le n° 7-12 de la revue "Studies in Family Planning" (revue du Population Council) sous la traduction de Jeanne Cairus Sinquefield.;

1. Utilisation

SPSS (Statistical Package for Social Sciences) est un système intégré de programmes pour l'analyse des données en sciences sociales. Le système a été conçu pour aider les chercheurs à réaliser d'une façon simple et très complète plusieurs types d'analyses de données à partir d'un logiciel général unique et au champ étendu. SPSS permet une assez grande souplesse dans le format des données. Il offre à l'utilisateur un vaste choix de procédures pour la transformation des données et la manipulation des fichiers, et propose au chercheur un grand nombre de programmes de calculs statistiques couramment utilisés en sciences sociales.

Outre les calculs usuels en statistique déscriptive (histogrammes, tableaux croisés, etc...), SPSS permet d'effectuer des calculs de corrélation simple, partielle ou multiple ou des calculs d'analyse factorielle des données. Les procédures de gestion des données permettent de modifier le dessin d'un fichier en même temps qu'on utilise les procédures statistiques. Ces possibilités permettent par exemple à l'utilisateur de générer de nouvelles variables, d'en recodifier d'autres, de sélectionner par sondage ou par choix raisonné des enregistrements, d'ajouter ou de modifier des données. SPSS permet au chercheur de conduire son analyse à partir de spécifications dans un langage naturel sans qu'il ait besoin de connaissances particulières en informatique. Le guide d'utilisation est écrit de telle façon que SPSS est accessible à un débutant en informatique.

.../...

2. Exigences techniques

Une version de taille modeste, SPSSG, est également disponible. Ce logiciel est écrit en FORTRAN IV pour un compilateur IBM de niveau G et nécessite environ 100 K. Attention à acquérir une version compatible avec le matériel. SPSS est disponible sur IBM 360 ou 370, sous OS ou DOS, CDC 3300 ou 600, CYBER 70, ICL 1900 ou 4130, BURROUGHS 4700, UNIVAC 1100 et XEROX. Il le sera prochainement sur BURROUGHS 3500, UNIVAC 1108 et NCR 201.

3. Evaluation des manuels d'utilisation et des programmes

Le manuel SPSS est un des meilleurs du genre. On y trouve beaucoup d'exemples, il ne nécessite pas la connaissance d'un langage de programmation de même que des spécifications du langage de commande.

L'ensemble des programmes qui constituent SPSS génèrent l'ensemble correspondant de procédures et ont en commun les conventions qui permettent de manipuler les données. Ils permettent à l'utilisateur d'exécuter une suite de tâches avec le minimum d'interventions manuelles de manipulation des données, etc... Les cartes dé définition des données doivent être préparées et entrées une seule fois et l'information qu'elles contiennent est sauvegardée avec les données correspondantes pour une utilisation ultérieure (l'utilisateur retrouve ainsi automatiquement les fichiers, les noms des variables, les formats, les labels, etc...). Outre cet avantage décisif, la vitesse de calcul est assez rapide car les fichiers sont traités en langage machine.

4. Avantages et inconvénients

Si on dispose de la taille-mémoire suffisante (100 K) et d'une version compatible des programmes, SPSS peut être acheté et installé. Il est très simple à utiliser et permet au chercheur des manipulations très fréquentes avec peu de travail de préparation quand on utilise une bande système. Son principal inconvénient est la taille mémoire nécessaire qui est importante. L'autre inconvénient est de n'autoriser le traitement que d'enregistrements de longueur fixe : il convient donc de donner aux enregistrements à traiter la même longueur en y ajoutant des colonnes ou des cartes blanches.

/

5. Comment se procurer manuel et programmes ?

Le manuel, écrit par Norman Nie et alii, peut être obtenu auprès de Mc-Graw-Hill Book Company. Il coûte 14,95 $ U.S. Un autre ouvrage, SPSS-Primer par William KELECKA, qui est un manuel d'enseignement pour étudiants, est disponible chez le même éditeur au prix de 4,95 $ U.S. Je recommande aux personnes intéressées l'achat d'un manuel avant la commande du logiciel général proprement dit.

Les programmes peuvent être commandés à l'adresse suivante :

S P S S Inc.
National Opinion Research Center
6030 Ellis Avenue
CHICAGO (Ill.) 60637

On vous fournira directement les programmes ou on vous orientera vers un centre de calcul qui dispose d'une version compatible. Son prix de vente est assez élevé : 750 $ U.S. pour les universités ou organismes de recherche, 1250 $ U.S. pour les associations sans but lucratif et 5000 $ U.S. pour les entreprises. Ce prix inclut 3 exemplaires du manuel. On peut aussi, sans que ce soit obligatoire, passer un contrat de maintenance au-delà de la première année de garantie, au prix annuel de 400 $ (universités ou centres de recherche) 600 $ (associations sans but lucratif) ou 2000 $ (entreprises). SPSS Inc. envoie gratuitement un bulletin d'information "SPSS Newsletter".

A N N E X E III

Présentation des logiciels généraux CENTS, CENTS-AID, et CO-CENTS

Avertissement : Ce texte est une traduction d'un article paru en anglais dans le n° 7-12 de la revue "Studies on Family Planning" (revue du Population Council) sous la signature de Jeanne Cairus Sinquefield.

1. Utilisations

CENTS (Census Tabulations Systems) est un logiciel de tabulation des données issues des recensements de la population et des habitations. Il est conçu pour effectuer des tableaux croisés et permet de préparer ces tableaux pour un grand nombre de zones géographiques différentes, de cumuler ces tableaux pour des zones plus vastes (publications) et de les éditer.

CENTS se compose de cinq programmes distincts, deux d'entre eux étant des utilitaires IBM de tri. Le premier programme, CENTAL, créé les cases des tableaux croisés selon les spécifications indiquées par le programmeur. Le second programme, TALSORT, est un programme utilitaire standard de tri. Il n'est utilisé que lorsque CENTAL crée plus de 591 cases de tableaux ou lorsque de nombreux niveaux géographiques exigent des cumuls. Le troisième programme, CENCON, utilise les sorties triées de TALSORT pour créer les tableaux cumulés prêts pour l'édition. Le quatrième programme, CONSORT, trie les tableaux selon les zones géographiques et prépare les noms en clair de ces tableaux pour l'édition finale. Le cinquième programme, CENPREP, génère les tableaux sous leur forme définitive.

CENTS résout deux gros problèmes qui se posent souvent lorsqu'on veut réaliser des tableaux croisés à partir d'un fichier d'unités statistiques : la taille du fichier et la structure des enregistrements. Si un fichier contient un grand nombre d'unités, comme c'est le cas pour les recensements, la production de tableaux croisés peut rapidement devenir très coûteuse. De plus, si le fichier est organisé selon une structure arborescente (par exemple, des logements suivis d'un nombre variable d'individus), et non selon une structure "à plat" (enregistrements de longueur fixe), le travail à réaliser n'est plus très simple. Beaucoup de logiciels standards de tabulation (comme SPSS) ne traitent que des fichiers "à plat" et ne permettent donc pas de fabriquer

.../...

des tableaux croisés qui relient des variables d'un enregistrement d'un certain type avec des variables d'un autre type d'enregistrement (par exemple caractéristiques du logement croisées avec caractéristiques individuelles). En conséquence, les fichiers hiérarchisés doivent être reformattés pour obtenir des fichiers "à plat" susceptibles d'être traités par ce type de logiciels ou bien il convient de résoudre de façon particulière chaque problème de tabulation en écrivant un logiciel spécifique. C'est pour ce genre de problème qu'a été conçu CENTS.

Toutefois, CENTS, bien qu'étant très efficace pour traiter des fichiers de structure hiérarchisée, nécessite un nombre impressionnant de cartes de contrôl Pour résoudre ce problème a été écrit le logiciel CENTS-AID qui intègre les cinq programmes constituant CENTS en un système unifié. Au lieu d'écrire et de tester "à la main" les programmes dans le langage de base (nécessaire pour CENTAL et CENTREP) et d'écrire les instructions détaillées pour TALSORT, CENCON et CONSORT, l'utilisateur n'a qu'à donner les spécifications dans un langage simple et naturel pour un seul système : CENTSAID. C'est ensuite CENTS-AID qui analyse ces commandes et les traduit en instructions pour CENTS, prépare l'exécution dans la séquence requise des programmes CENTS et génère les tableaux demandés. Le but de CENTS-AID est de combiner l'efficacité exceptionnelle de CENTS avec les spécifications simples que requiert l'utilisation de logiciels tels que SPSS. Le résultat est un système très rapide de tabulation à la portée d'un utilisateur n'ayant qu'un minimum de connaissances informatiques grâce à un langage de command très naturel.

2. Exigences techniques

Le logiciel CENTS de base est disponible en quatre versions :

a) le logiciel d'origine est écrit en assembleur IBM (ALC : Assembly Language Coding), utilise un compilateur de niveau F sur un IBM 360 ou 370 sous OS, DOS ou TOS et nécessite une mémoire de 24 K pour fonctionner, bien qu'il tire partie de toute addition de mémoire.

b) COBCENTS est une version du CENTS écrite en COBOL (ANS/niveau 2) indépendante du matériel : il nécessite 54 K de mémoire sur un ordinateur disposant d'un compilateur COBOL. Il tire également partie de tout accroissement de la taille-mémoire en permettant l'augmentation du nombre de tableaux réalisés en un seul passage.

.../...

c) CENTS-AID est, sous sa forme d'origine, écrit en assembleur IBM (ALC) de niveau F, et nécessite un IBM 360 ou 370 sous OS ou DOS d'au moins 42 K. Sur option CENTS DOC permet de générer un fichier documentaire et une documentation écrite sur les fichiers "entrée"de CENTS AID ; il est écrit en COBOL niveau 2 et nécessite 54 K.

d) CENTS-AID II est écrit en COBOL niveau 2 et nécessite 80 K au moins. Il peut fonctionner sur IBM 360 ou 370 sous DOS ou OS, sur les ordinateurs CDC de la série 6000 et sur XEROX sigma 5.

3. Evaluation des Manuels d'utilisation et des logiciels

Les manuels pour CENTS et CO-CENTS sont très difficiles à assimiler par des personnes n'ayant pas un minimum de connaissances informatiques. Deux des cinq programmes utilisent un langage de programmation d'un haut niveau technique, et les trois autres nécessitent un grand nombre de cartes de contrôle. Toutefois, quand il faut tabuler un grand nombre de données, particulièrement quand la structure des fichiers est hétérogène, comme dans les recensements, ces logiciels sont remarquablement efficaces. L'utilisateur aura avantage à se faire aider par un technicien expérimenté pour les utiliser au mieux de leur efficacité.

CENTS-AID, en revanche, s'accompagne d'un manuel agréablement écrit, quoique manquant d'exemples précis. Il ne requiert pas la connaissance d'un langage, mais nécessite toutefois des connaissances sur la structure des fichiers informatiques à tabuler. CENTS-AID peut être utilisé sans assistance extérieure par quelqu'un n'ayant qu'un minimum d'expérience en informatique et peut aisément être enseigné à des débutants. Tous ces logiciels peuvent facilement être installés par un technicien expérimenté.

4. Avantages et inconvénients

Les avantages de ces trois logiciels sont leur efficacité (comparativement à leur coût) pour la tabulation de gros fichiers et leur capacité à traiter les fichiers arborescents souvent utilisés dans les recensements. Ils sont vivement recommandés aux bureaux de recensement ne disposant que d'ordinateurs petits ou moyens. CENTS-AID est préférable à CENTS ou CO-CENTS du fait de sa maniabilité. Toutefois, la taille mémoire et le type de l'ordinateur disponible peuvent nécessiter l'utilisation de CENTS ou de CO-CENTS. Pour le statisticien ces logiciels ont l'inconvénient de ne pas permettre le calcul d'indicateurs statistiques à l'exception des pourcentages ; seul CENTS-AID II permet le calcul

/

d'un nombre limité d'indicateurs (pourcentages, moyenne, médiane, variance et χ^2) Pour des études plus modestes (moins de 5000 cases ou fichiers "à plat") ou lorsque le calcul d'indicateurs statistiques plus sophistiqués est désiré, d'autre logiciels sont recommandés (SPSS, par exemple).

5. Comment se procurer les manuels et les programmes

Les manuels et les programmes CENTS et CO-CENTS peuvent être obtenus gratuitement à l'adresse suivante :

Computer Methods Library,
International Statistics Programm Center
U.S. BUREAU OF THE CENSUS
WASHINGTON (D.C.) U.S.A.

Le Bureau of the Census a, dans le passé, apporté son assistance pour installer ces programmes.

Le programme CENTS-AID (version 1) sous OS ou DOS coûte 100 $ U.S. Le programme CENTS-AID II qui permet les calculs de moyenne, médiane, variance et χ^2 coûte 650 $ U.S., y compris une année de maintenance. Les commandes doivent être expédiées à :

D U A Labs. (DATA USE and ACCES LABORATOIRES)
Suite 900
1601 Kent Street
ARLINGTON (Virginia) 22209, U.S.A.

L'envoi des programmes s'accompagne du manuel approprié. Il convient de spécifier le nombre de pistes, la densité et les options label de bande lors de la commande (7 ou 9 pistes ; 556, 800 ou 1600 bpi, labels standard OS ou sans label). Il est admis que le manuel puisse être acheté seul à un coût plus bas avant l'achat des programmes.