P A N E L S O C I O - E C O N O M I Q U E

" L I E W E N Z U L E T Z E B U E R G "

Document PSELL No. 9

L O G I S T I Q U E

&

D O C U M E N T A T I O N

Principes d'organisation de la documentation dans le panel

J. Tournois

Document produit par le

CENTRE D'ETUDES DE POPULATIONS; DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES
a.s.b.l.

B.P. 65 L-7201 Walferdange
Tél. (352) 33 25 15

Président: Gaston Schaber

---

1 9 8 8

SOMMAIRE

INTRODUCTION

En 1985, naissait le Panel Socio-économique "Liewen zu Letzebuerg". A cette date, l'organisation du panel était conceptualisée par un schéma très simple sous deux rubriques essentielles : PRODUCTION et ANALYSE.

Les tâches de production consistaient en production des questionnaires, collecte des informations, réception et contrôle des enquêtes, codage des informations et encodage des données en vue de l'élaboration d'un fichier informatique analysable.

A partir de ce (ces) fichier(s) de départ, l'analyste entrait en action, partageant son temps entre les contrôles et corrections de ce(s) fichier(s), la création de variables nouvelles plus propres à être analysées, la constitution de nouveaux fichiers et, finalement, l'analyse proprement dite.

Devant la lourdeur de ces tâches préparatoires à l'analyse, en particulier celle des contrôles et corrections, une première amélioration consista à rendre plus efficace l'encodage des données de manière à minimiser les erreurs.

Alors que les enquêtes de 1985 étaient encodées manuellement et les erreurs corrigées après-coup, l'encodage des données de 1986 passe par l'intermédiaire d'un programme de saisie des données. Ce programme, d'une part, rend plus conviviale la tâche d'encodage ce qui produit une diminution des erreurs; d'autre part, il offre toute une série de vérifications logiques détectant les erreurs dans le temps même de l'encodage et permettant leur rectification immédiate.

La réalisation de ce programme appartenait déjà au service "logistique et documentation" avant même que celui-ci ne soit créé.

L'expérience de la vague 1985 suggère aussi que d'autres dispositions soient prises.

En 1985, les 620 questions du questionnaire ne conduisent pas à 620 variables mais à 1900 variables réparties en 3 fichiers de travail correspondant aux trois niveaux d'analyse (l'individu, le groupe de revenu, le ménage).

De plus, aucun de ces fichiers n'est créé directement sur la base des définitions des variables. Par exemple, il a fallu passer par l'intermédiaire de 17 fichiers avant d'aboutir au fichier actuel de travail sur les ménages.

Devant cette masse énorme d'informations, la **nécessité d'une documentation** suivie s'est très vite fait sentir. Dès mars 1987, P. HAUSMAN, dans une communication aux Premières Journées Nancéiennes sur l'Analyse Dynamique de la Pauvreté (*: communication non publiée), insistait sur l'importance de cet aspect documentaire dans toute étude par panel.

Cette documentation doit notamment renseigner sur les variables, les fichiers dans lesquels ces variables sont disponibles et sur l'emplacement physique de ces fichiers informatiques.

Enfin, une étude par panel est par définition une étude longitudinale, ou dynamique, dans laquelle les sujets sont interrogés plusieurs fois de suite. Une implication immédiate est la nécessité d'un suivi des individus qui composent l'échantillon, de manière à pouvoir les retrouver et les interroger à nouveau lors des vagues successives.

Une autre implication est la nécessité de **réunion des données des vagues successives** en un seul fichier de travail. Dans le cas présent, il s'agit de se donner les moyens de réunir, dans de bonnes conditions, les informations des deux premières vagues du panel (l'année 1985 et l'année 1986).

La prise en compte dans un fichier unique d'informations en provenance de différents fichiers suppose que des variables spécifiques soient créées. Toutes les variables servant la liaison des différentes informations peuvent être réunies dans un **fichier de gestion de l'information longitudinale** qui doit permettre, à terme, pour chaque étude entreprise, la sélection des individus (groupes ou ménages) pour lesquels cette étude a un sens et des variables pertinentes pour l'analyse.

La constitution de ce fichier de gestion longitudinale est une tâche aussi délicate que cruciale. Elle suppose une parfaite cohérence des informations (nécessitant des contrôles) et une harmonisation des variables.

L'élaboration de ce fichier ainsi que les différents travaux qui en constituent les préliminaires appartiennent à la partie **logistique**.

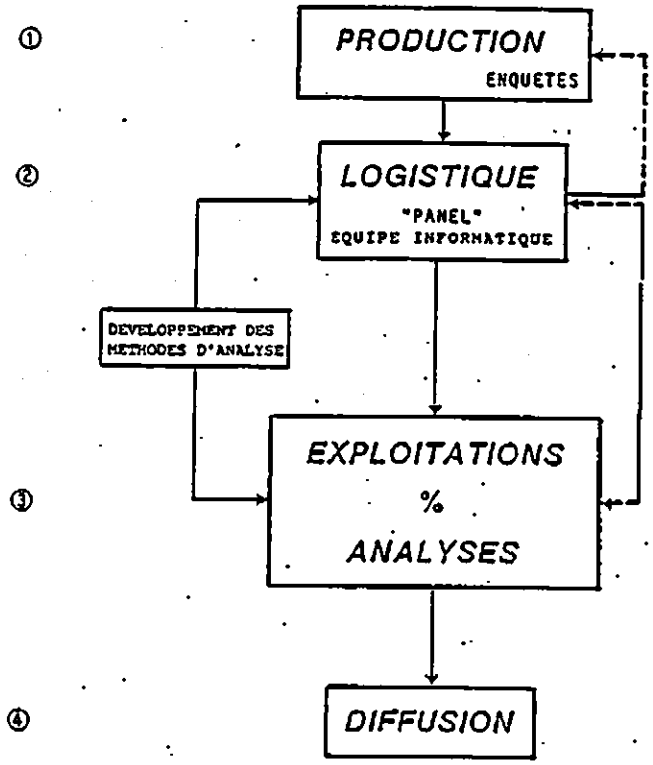
Ainsi, autour des différentes tâches d'organisation, de gestion et de maintenance de la masse croissante d'informations que fournit une étude par panel, se constitue peu à peu un véritable service que nous dénommons "logistique et documentation".

Quoique ce service puisse intervenir à tout moment du processus d'élaboration d'une recherche, (depuis la collecte jusqu'à la diffusion en passant par l'analyse), sa place logique le situe **entre la production et l'analyse**.

---

*) A.D.E.P.S., C.E.P.S., I.N.S.E.E. - Premiers jalons pour une analyse dynamique de la Pauvreté / précarité. Cahiers économiques de Nancy, n.18, 1987.

ORGANISATION GENERALE
DES ACTIVITES DE LA RECHERCHE



Il est possible de rassembler les différents travaux à accomplir à ce niveau sous deux rubriques qui se complètent : LOGISTIQUE et DOCUMENTATION.

La partie logistique regroupe toutes les tâches de gestion, d'organisation et de maintenance de l'information destinée à l'analyse.

Quant à la partie documentation, elle vise à mettre à la disposition de l'analyste toute information utile sur la nature de l'information, sa localisation et sa disponibilité.

Il est clair, d'une part, que l'élément logistique ne peut être mené à bien sans une bonne connaissance préalable des différents niveaux de l'information (niveau du questionnaire, niveau du codage et de l'encodage, niveau des fichiers d'analyse) et, d'autre part, qu'une documentation solide à chacun de ces niveaux ne peut être établie sans une visée d'ensemble du processus même d'élaboration de cette information et sans la prise en compte des liens entre ces différents niveaux.

Le présent document se propose, à partir des travaux réalisés au cours de l'année 1987 dans le cadre de la documentation et de la logistique attachées au Panel Socio-économique luxembourgeois, d'exposer les options qui ont été prises dans l'élaboration de l'information à analyser, les raisons qui ont présidé à ces choix, les difficultés rencontrées et les moyens mis en oeuvre pour apporter des solutions.

Compte tenu des avantages incontestés que présente une étude par panel du point de vue de la qualité des recherches qui en découlent, il est fort probable que ce genre d'étude ne soit peu pratiqué qu'en raison de la lourdeur et de la difficulté des tâches de gestion qui s'y rapportent. Une conséquence immédiate est le peu de documentation sur ce type de travail.

L'objectif de ce document est donc de mettre à la disposition du chercheur que l'étude par panel intéresse, le fruit d'une expérience nouvelle.

Un premier chapitre sera consacré à situer plus spécialement les travaux logistiques et documentaires dans le processus global d'élaboration de la recherche. Il donnera quelques repères et quelques points-clés ainsi que des principes généraux de fonctionnement. A titre d'illustration seront présentées les grandes lignes du programme de saisie de données.

Le second chapitre sera consacré aux travaux de documentation, en particulier à la documentation de nature transversale (c'est à dire celle qui se rapporte à une vague). Des éléments documentaires réalisés à partir de traitement de texte et base de données seront présentés à titre illustratif et comme support de réflexion.

Au cours du troisième chapitre, nous serons ainsi "armés" pour aborder une perspective plus proprement logistique. Celle-ci traitera plus spécialement des pré-requis à la réalisation d'un fichier de gestion longitudinale et sera principalement illustrée par le système de nomenclature longitudinale des variables.

Le dernier chapitre sera essentiellement prospectif. Après un bilan des travaux effectués, nous estimerons ceux qui restent à accomplir pour compléter la réalisation entreprise, et proposerons les éventuels aménagements pour son amélioration et des éléments de réflexion pour sa poursuite.

Nous verrons à cette occasion que ces points se présentent, à long terme, comme des estimations des réalisations nécessaires à la mise en place d'un quasi "système-expert" pour les travaux d'étude par panel.

I - SITUATION DES TRAVAUX LOGISTIQUES ET DOCUMENTAIRES

1.1 - La place des travaux "logistique et documentation" dans le processus de recherche.

La place logique du service "logistique et documentation" se situe entre la PRODUCTION et l'ANALYSE.

En effet, la documentation s'établit nécessairement après que les observations aient été récoltées.

D'autre part, le service logistique entre en jeu dès que ces observations sont recueillies. Sa fonction est de produire les données qui seront prêtes pour l'analyse et de documenter tous les aspects pertinents pour l'analyste. On peut ainsi envisager l'aspect logistique comme un énorme processus de transformation des observations en données.

Pour en rendre compte, nous empruntons la terminologie de COOMBS (1). Selon cet auteur, l'analyste ne travaille jamais sur des observations, mais sur des données.

Une observation est l'information brute telle qu'elle est récoltée.

Une donnée est une information transformée en vue d'être analysée.

COOMBS explique qu'il y a toujours un processus de transformation des observations en données avant qu'une analyse ne soit pratiquée. Cette transformation est de nature logique ou mathématique. Sans entrer dans le détail de ces transformations, disons qu'en ce qui nous concerne directement, les observations seront codées, puis organisées en une matrice rectangulaire (un fichier informatique).

Ces transformations peuvent prendre une forme minimale; mais, même dans ce cas, elles deviennent rapidement très difficiles à gérer tant le nombre d'informations croît.

---

(1) - COOMBS C. - A theory of data , Mathesis Press, Ann Arbor, 1976.

Dans une entreprise comme le panel luxembourgeois, ces transformations sont très importantes :

- en raison du nombre des observations de base (environ mille "observations" par individu dès le départ);
- du fait de l'organisation en trois niveaux d'analyse (par exemple, des données individuelles sont agrégées au niveau du ménage ou des données relatives aux ménages sont répercutées sur les individus membres d'un ménage);
- du fait de la dimension temporelle : de nombreux indicateurs doivent être construits pour permettre de rattacher longitudinalement l'information aux différents niveaux d'analyse.

Ce processus d'élaboration des données en fichiers analysables, une fois l'information de base recueillie, chemine par les étapes suivantes:

- VERIFICATION rapide dès réception de l'ENQUETE.

- CODAGE et vérification approfondie des ENQUETES.
  - codification des données manquantes.
  - correction des anomalies rectifiables.

- ENCODAGE et détection des codes sauvages; soit manuellement (année 1985), soit par programme (année 1986)

- détection et CORRECTION DES ERREURS LOGIQUES suivant un processus en trois temps)
  - fabrication de fichiers provisoires.
  - rédaction et exécution de programmes de détection des erreurs logiques.
  - rectification des erreurs logiques.

- CONSTITUTION DE FICHIERS ELEMENTAIRES.

- CREATION DES VARIABLES DE TRAVAIL qui s'ajoutent aux variables élémentaires.

- CONSTITUTION et stockage DES FICHIERS DE TRAVAIL.

Ces grandes étapes concernent tant une étude transversale qu'une étude longitudinale. A ce stade, l'aspect longitudinal n'intervient qu'indirectement :

- en augmentant le nombre des variables; certaines variables ne servent qu'à établir la liaison inter-vagues,

- en entraînant la constitution d'un fichier de gestion longitudinale, qui n'est pas un fichier de travail au sens propre mais un fichier de sélection des informations sur lesquelles une analyse longitudinale peut être effectuée.

- en augmentant la complexité des vérifications et corrections : des contrôles de cohérence de l'information longitudinale doivent être entrepris (par exemple, une "femme" en 1985 ne peut être devenue un "homme" en 1986);

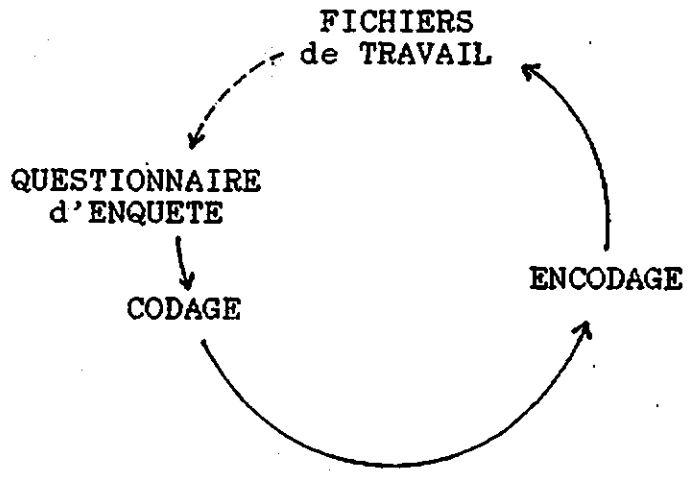
- en rendant crucial l'aspect documentaire; l'analyste doit être renseigné sur l'évolution des données qu'il traite, au fil des vagues successives.

L'aspect longitudinal a une autre incidence générale.

Dans le déroulement linéaire du processus de transformation des observations en données, depuis le formulaire d'enquête jusqu'aux fichiers de travail, le processus doit se poursuivre l'année suivante.
Sa continuation immédiate réside dans le questionnaire d'enquête de la vague suivante.
On peut le schématiser comme suit :

```
                                                          A
                                                          N
                                                          A
                                                          L
ENQUETE —> CODAGE —> ENCODAGE —> FICHIERS —————————Y———> ENQUETE
 temps                                                    S   temps
   t                                                      E    t+1
                                                          S
```

Au total, comme l'enquête par panel renouvelle au temps t+1 le questionnaire du temps t, on peut considérer qu'à chaque vague une boucle est constituée.
Et le schéma de fonctionnement de l'élaboration de l'information est plus justement représenté par un processus circulaire.



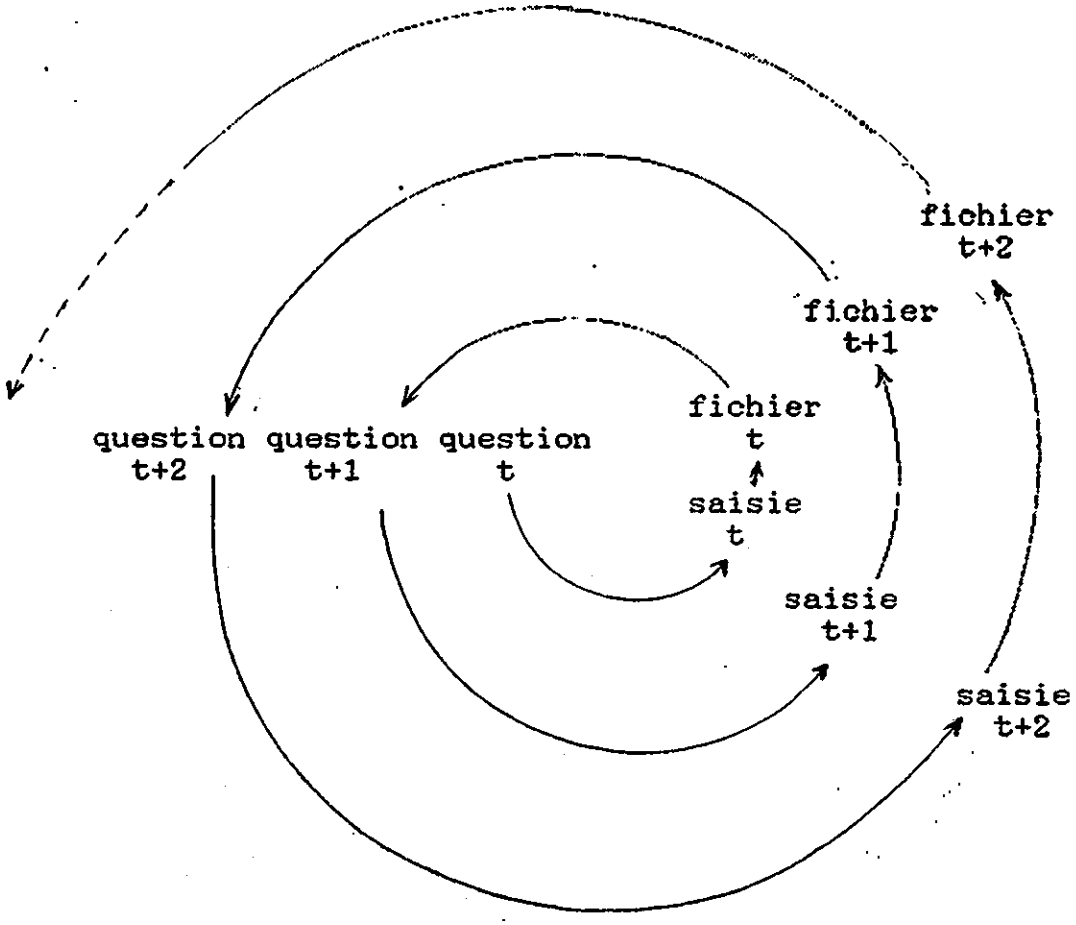
Mais, quoique par nature le questionnaire des vagues successives doive comporter les mêmes questions pour constituer un panel, l'expérience montre que l'enquête au temps t+1 n'est pas absolument la réplique exacte de l'enquête au temps t ... ne serait-ce que parce que des erreurs doivent être rectifiées.

A tout moment du processus, on peut diagnostiquer des imperfections dans le questionnement de base, y compris lors des phases ultimes d'analyse.

Des intérêts nouveaux peuvent se faire jour et/ou des pistes de recherche se révéler stériles dès les premières analyses.

En bref, le questionnaire, quoiqu'il conserve un "noyau dur" de questions qui ne changent pas, subit au fil du temps certaines modifications. Et il y a un certain décalage entre les questionnaires des vagues successives, traduisant une évolution.

Au total, l'évolution d'un panel est un processus en spirale.



L'élaboration de l'information poursuit ainsi un cheminement continu et des décisions qui sont prises à un moment quelconque ont des répercussions à long terme.

La position occupée par le service logistique a bien évidemment une incidence directe sur les analyses, par le biais des fichiers de données qu'il constitue.

Il a aussi une incidence indirecte. Toute l'information passe par ce service, à tous les stades de son élaboration. Il est ainsi à même de remarquer, à un moment quelconque de ce cheminement, des imperfections à redresser; de ce fait, à solliciter que des aménagements soient pris, soit en amont du processus logique pour l'élaboration des fichiers d'analyse, soit en aval pour la rédaction du questionnaire de l'année suivante.

Il ne paraît pas utile d'insister sur la situation-clé qu'occupent les tâches de logistiques dans l'élaboration d'un panel.

1.2 - Les principes de fonctionnement.

Il est clair que les travaux logistiques sont au service de l'analyse puisque la fonction première de ce service est de constituer les fichiers de données à analyser.

Nous venons de noter que ce service manipule de l'information, puisque sa tâche peut se résumer par la formule: il transforme les observations en données.

D'autre part, nous avons signalé un long cheminement dans l'élaboration des informations, depuis le questionnaire de départ jusqu'aux fichiers finaux.

Il est aussi manifeste que les fichiers sur lesquels l'analyste travaille doivent refléter aussi fidèlement que possible les observations de départ. L'analyste, à la phase ultime de ce long cheminement de l'information sous ses diverses formes, ne doit pas être coupé de la réalité sur laquelle il travaille : les réponses que les sujets interrogés ont fournies à l'enquêteur; ces réponses, telles qu'elles sont transcrites sur le questionnaire.

En d'autres termes, la tâche du service logistique est de raccourcir le chemin logique de l'élaboration de l'information. L'analyste ne doit pas perdre de vue le questionnaire de départ, même si - dans la réalité - les informations que ce dernier contient, se présentent sous diverses formes successives.

Il ne s'agit en aucune façon de brûler les étapes de l'élaboration de l'information. Elles sont nécessaires.

Le seul remède à ce fossé qui tend à se creuser entre l'information de départ et l'information finale est le principe d'homogénéisation de l'information.

Une illustration en sera donnée ultérieurement à travers la nomenclature des variables. Pour le moment, éclairons simplement la conception de base à travers une seule question du questionnaire.

Question: "la personne a-t-elle des difficultés à parler le français ?"

c'est la question	B24	du questionnaire B en 85
c'est la variable	V525	du fichier d'analyse de 85.
c'est la question	C8	du questionnaire C de 1986 dans le cas d'un enfant.
c'est la question	D7	du questionnaire D de 1986 dans le cas d'un adulte.
c'est la variable	C81	de la saisie de 1986 dans le cas d'un enfant.
c'est la variable	D71	de la saisie de 1986 dans le cas d'un adulte.
c'est la question	C3	du questionnaire 1987 dans le cas d'un enfant.
c'est la question	D20	du questionnaire 1987 dans le cas d'un adulte.

Il est clair que si l'analyste veut retourner à l'information de base, ou s'il a besoin de suivre son évolution, il devra disposer d'une table de correspondance qui lui signifie que

V525 = B24 = C8 = D7 = C81 = D71 = C3 = D20.

Et si l'on songe qu'il y a plus de mille variables en jeu sans compter les variables créées, il est clair qu'on ne peut pas faire un travail correct dans ces conditions.

L'application du principe d'homogénéisation, dans ce cas, consiste simplement à harmoniser les désignations des informations.

Pour l'analyste, la V525 de 1985, disons la V85525, doit rester la V525 dans le questionnaire, lors de la saisie, dans les fichiers de stockage et dans les fichiers de travail.
Et idéalement, on doit arriver à la solution où la désignation reste stable en ne mentionnant que les changements d'année, à savoir

V85525 = V86525 = V87525 = etc.

On voit que le principe d'homogénéisation n'est souvent qu'un corollaire du principe de simplicité.

On s'accordera facilement sur le fait qu'un système simple est plus facile à manipuler qu'un système complexe et de ce fait entraîne moins d'erreurs.

Dans le même ordre d'idées, on convient que les tâches fastidieuses entraînent de nombreuses erreurs. Il en va ainsi des tâches de codage et d'encodage : l'expérience de la vague 1985 a montré de très nombreuses erreurs d'encodage nécessitant un temps de correction très long.

Aussi, un des objectifs de l'investigation du service logistique est de tenter de minimiser les tâches fastidieuses et répétitives en vue de minimiser les erreurs.

Le programme de saisie des données, réalisé pour faciliter l'encodage de la vague 1986, est une illustration de ce principe.

Quoique la rédaction d'un tel programme soit complexe, son principe de fonctionnement en est simple. Notre intention n'est pas d'entrer dans le détail de sa présentation, mais seulement d'exposer les grandes lignes de son fonctionnement.

Alors que l'encodage traditionnel consiste à transcrire dans la machine des pages entières de chiffres totalement hermétiques, l'encodage qui est gouverné par ce programme, présente à l'écran des pages qui rappellent tout à fait les pages du questionnaire (les "panneaux") : l'encodeur écrit, dans les emplacements prévus à cet effet, les codes qu'il lit directement (au même emplacement) sur le questionnaire. Le code qu'il transcrit est devenu signifiant.

Par ce simple fait, par l'aspect convivial de la présentation de ce type d'encodage, le nombre des erreurs de transcription est diminué.

Il est en effet peu probable, par exemple, que quelqu'un écrive 1996 en face de la ligne "année de naissance", si ce n'est par erreur de frappe.

D'autre part, chaque panneau est assorti, lors de sa fabrication, d'un programme de vérification des erreurs d'édition ou de logique, qui avertit l'encodeur des erreurs détectées, à mesure qu'elles apparaissent, au moyen d'un message. L'encodeur peut immédiatement les redresser.

Par exemple, l'encodeur a frappé par mégarde 1996 pour une année de naissance; le programme est conçu de telle façon qu'aucun chiffre non compris entre 1880 et 1986 ne soit accepté à cet endroit. Il émet donc un message à l'écran explicitant :

"L'année de naissance ne peut pas
être supérieure à 1986."

L'encodeur peut ainsi redresser immédiatement son erreur.

Les autres dispositions qui régissent le fonctionnement du service logistique découlent des qualités mêmes de l'information.
Pour qu'une information soit utile, elle doit être claire, univoque, facilement accessible.

De là découle la nécessité d'une documentation précise à l'usage de l'analyste.

La masse énorme d'informations contenues dans un panel comme celui-ci n'a de sens que par rapport aux analyses qu'elle permet de réaliser. Et ces analyses ne se conçoivent pas sans un système de documentation sur la nature des données disponibles.

Pour que l'information soit porteuse de fruits, elle doit être rendue accessible :

- soit à des nouveaux membres de l'équipe de base,
- soit à des personnes extérieures à l'équipe, autorisées à réaliser certaines analyses,
- soit à des évaluateurs externes.

A long terme, dans une enquête longitudinale où la masse d'information est nécessairement croissante, l'absence d'harmonisation et d'accessibilité de l'information ne pourrait qu'être stérile.

1.3 - l'enjeu d'une gestion de l'information.

On aura vite compris que la quantité même de l'information nécessite sa gestion concertée.

Pour comprendre l'enjeu d'une telle entreprise, il faut aussi prendre en compte le fait que la gestion d'un panel est un système dynamique.
Le temps de la gestion est rarement le présent. Le panel est délibérément tourné vers l'avenir. Il est aussi relié au passé et travaille sur de l'information mouvante.

Pour exposer ces points, présentons la situation au 31-12-1987.

ANNEE 1985

Le processus complet de transformation des observations en données est achevé.
Les données de 1985 sont en permanence analysées.

ANNEE 1986

Les observations de 1986 ont été récoltées, codées, encodées par le programme de saisie, stockées dans le système ISQL.
Ces informations subissent les derniers contrôles de cohérence qui les conduisent aux données des premiers fichiers élémentaires.

ANNEE 1987

Les observations de 1987 ont été récoltées et vérifiées de façon approfondie.
Le nouveau programme de saisie de données de 1987 est en cours de réalisation.

ANNEE 1988

Le questionnaire de 1988 est en cours de réalisation.

La situation au 31-12-1987 est donc celle-ci : avant même que ne se pose la tâche spécifique de jonction de deux vagues d'enquête, le travail effectif se situe sur quatre vagues.

On comprend sans peine qu'une décision prise au moment T a des implications au moment T+1 et que les tâches de gestion doivent être essentiellement prospectives.

Il faut ajouter que cette décision du temps T devra être répercutée au temps T-1 si l'on ne veut pas perdre la cohérence longitudinale des informations. Or, cette cohérence se présente comme un impératif du point de vue de l'analyste.
Les tâches de gestion deviennent donc aussi rétrospectives.

Dans la pratique, la décision de systématiser la nomenclature des variables de manière à permetre les analyses longitudinales, implique :

- que la saisie 1987 soit réalisée avec les nouveaux noms;

- que les variables 1986 soient renommées en cours de travail;

- que le questionnaire 1988 prévoie ces nouveaux noms avant son impression;

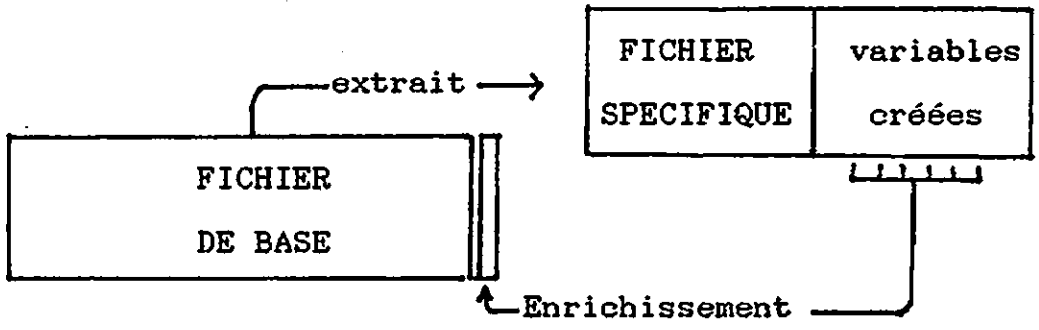
- que toutes les variables de l'année 1985 (dans tous les fichiers) soient renommées;

- que la documentation soit mise à jour.

En d'autres termes, la constitution des fichiers 1985 aura été faite deux fois.
Le travail sur les années 1986 et 1987 doit être effectué à l'aide de tables de correspondance.
Seul, le travail à venir concernant l'année 1988 sera facilité.

Outre le double mouvement prospectif et rétrospectif, il convient de signaler que l'information est en perpétuel mouvement.
L'information en cours d'élaboration est évidemment une information qui se transforme.
Mais quand le stade de la transformation est achevé, les données continuent d'être analysées et ces analyses génèrent de nouvelles données qui modifient les fichiers de travail.

Par exemple, un extrait d'un fichier de travail est réalisé pour une recherche spécifique. A cette occasion, des variables sont créées pour mener à bien les analyses projetées. Parmi ces variables, certaines présentent un intérêt d'ordre général. Elles seront reportées dans le fichier de travail de base, ce qui constitue un enrichissement pour les nouvelles analyses.



Ainsi, même lorsque les tâches de gestion se situent dans le présent, elles se situent dans un présent "mouvant".

Il devient donc impératif de pouvoir tenir à jour l'état des données dans les fichiers de travail, à mesure que celles-ci sont créées. Et cette mouvance constante des données ne fait elle-même que renforcer la nécessité d'une documentation.

Il ne paraît pas utile d'ajouter ici d'autres commentaires et le lecteur intéressé saura tirer les conclusions qui s'imposent.

L'expérience a montré que l'instauration d'une documentation précise et la pratique d'une gestion concertée de l'information doivent être mises en place dès le point de départ d'un panel (*).

---

*) : Faute de ressources suffisantes, ceci n'a pu être réalisé d'emblée dans le panel socio-économique luxembourgeois.

II - LA DOCUMENTATION.

2.1 - Quelques repères.

2.1.1 - Documentation transversale et documentation longitudinale.

Il est pratique de distinguer la documentation transversale de la documentation longitudinale.

- La documentation transversale est celle qui renseigne sur une vague du panel. Elle est, dans le panel luxembourgeois, établie au titre d'une année.
- La documentation longitudinale, comme son nom l'indique, concerne plusieurs vagues. C'est une documentation inter-vagues. Elle traite essentiellement des liaisons entre les vagues successives.

Par exemple, une liste des variables créées à partir des données de base de 1985 appartient à la documentation transversale. Elle ne concerne que la vague 1985. Pour l'année 1986, une autre liste devra être établie, qui ne concernera à nouveau que 1986.

Mais une liste des questions introduites en 1986 alors qu'elles n'existaient pas en 1985 fait partie des documents longitudinaux. Le lecteur est renseigné, par cette dernière, sur les informations qu'il ne pourra pas trouver lors du passage 1985-1986.

Dans ce chapitre, il sera exclusivement question de documentation transversale.

2.1.2 - Traitement de texte et base de données.

Il est clair que l'ère "papier-crayon" est révolue dans une telle entreprise. Quand nous parlons de documentation, nous parlons de documentation informatique.

Des documents sont toujours présents dans toutes les enquêtes:

- la bible de codification des variables qui accompagne le questionnaire,
- les dictionnaires de variables qui accompagnent les fichiers de données.

Mais ces documents traditionnels sont insuffisants.

En reprenant les termes de P.HAUSMAN, lors de sa

communication aux Pemières Journées Nancéiennes sur l'Analyse Dynamique de la Pauvreté (mars 1987), nous considérerons que pour gérer un panel,

- chaque information de base doit être identifiée, résumée, fichée;
- pour tout indicateur, il doit être possible de se faire facilement une idée quant à son mode de construction;
- chaque information doit être répertoriée par rapport au niveau d'analyse où elle est accessible;
- l'organigramme retraçant les voies de fabrication des fichiers doit être conservé, mis à jour et transcrit sous une forme compréhensible;
- l'histoire de chaque item appartenant au "corps dur" doit être constituée et stockée (et mentionner les éventuels changements de formulation et/ou de niveau d'analyse).

Ces éléments ne constituent que quelques exemples se rapportant à la documentation aussi bien transversale que longitudinale.

Pour aborder ces différents aspects, un outil informatique vient immédiatement à l'esprit: la base de données relationnelle.

Les logiciels de cette catégorie sont spécialement conçus pour gérer des données et établir des liens entre elles. Ils offrent l'avantage de permettre des tris multiples et des sélections de données sous condition logique. Les données se prêtent à des restructurations multiples, les différentes bases sont modulables; elles peuvent être liées entre elles. Des possibilités de programmation permettent de rédiger des macro-instructions si une série de demandes doit être répétitive. Elles permettent d'effectuer des impressions sur papier dans différents formats, ...

Leurs avantages sont multiples.

Par rapport à notre objectif, elles ne présentent qu'un inconvénient.

Pour que des éléments soient introduits dans une base de données, qui se présente comme une structure, elles doivent être structurées. Or, toutes les données dont nous disposons ne se laissent pas forcément emboîter dans une structure.

Le choix fut donc le suivant. Chaque fois qu'il a été possible de structurer l'information pour laquelle une documentation devait être entreprise, il a été fait recours à une base de données. Quant ce n'était pas possible, la documentation a été complétée par traitement de texte.

Les logiciels de traitement de texte présentent aussi des intérêts, ne serait-ce que parce qu'ils permettent de faciles mises à jour. De plus, dans un panel, la documentation établie pour une vague est facilement reprise comme point de départ pour établir celle de la vague suivante.

Pour éviter des pertes de temps, il est seulement nécessaire de recourir à des logiciels de base de données et

de traitement de texte compatibles, de manière à pouvoir utiliser l'information en provenance de la base de données dans le traitement de texte, sans avoir à la réécrire.

2.2 - Documentation transversale par traitement de texte.

Le traitement de texte a été essentiellement utilisé pour établir une documentation sur les variables nouvelles (ou variables calculées), c'est-à-dire les variables créées par programme à partir des variables de base qui proviennent du questionnaire.

Il s'agit dans ce cas de retracer l'histoire de la création de chaque variable. Le choix du traitement de texte repose sur la considération du fait que ces histoires sont très différentes. Quelquefois, il s'agit du simple recodage de variable élémentaire; dans d'autres cas, l'histoire est fort mouvementée et passe par l'intermédiaire de plusieurs fichiers, avec des agrégations, etc.

Il semble exclu de pouvoir prévoir "le cheminement le plus complexe possible" qui englobe en lui "tous les chemins possibles" afin de dégager une structure qui rendrait le document accessible à une base de données. De plus, la plupart des éléments de cette structure seraient vides.

Ainsi, trois documents sont réalisés. Ils correspondent aux trois fichiers de travail de la vague 1985, ces fichiers majeurs étant établis par niveau d'analyse :

- l'individu,
- le groupe de revenu,
- le ménage.

- V.N.I. retrace l'histoire des Variables Nouvelles Individuelles.

- V.N.G. retrace l'histoire des Variables Nouvelles du Groupe de revenu.

- V.N.M. (en cours de rédaction) concerne les Variables Nouvelles du Ménage.

La vocation essentielle de ces documents est d'éviter à l'analyste de devoir relire (déchiffrer) les programmes de création chaque fois qu'il a besoin de connaître la signification précise d'une variable et, pour ce faire, d'avoir à se reporter à de nombreux fichiers. Ces documents sont donc rédigés en langage clair; ce ne sont pas des copies des programmes de création.

Il est clair que le rédacteur de ces documents doit être lui-même suffisamment au courant de l'analyse de données pour sélectionner l'information qui est susceptible d'intéresser l'analyste.

Toutefois, les programmes de création des variables restent disponibles, en cas de besoin, soit sous forme informatique (s'ils doivent être réutilisés), soit sous forme imprimée.

Les indications pour cette consultation sont répertoriées dans une base de données. Nous verrons ce point ultérieurement (§ 2.3).

Ces documents doivent être tenus régulièrement à jour puisque des variables nouvelles peuvent toujours être créées par les analystes et se révéler d'un intérêt général. Dans ce cas, elles enrichiront les fichiers de travail et seront conservées dans le fichier (cf. § 1.3).

Il s'agit clairement de documents de travail, qui mentionnent au besoin diverses anomalies ou ambiguïtés rencontrées dans les programmes.

A terme, toutefois, une fois que les créations sont stabilisées, ils contribuent à l'élaboration d'une bibliothèque constituant la mémoire du panel.

Un de ces documents, le V.N.I., est présenté en illustration (page 22 et suivantes).

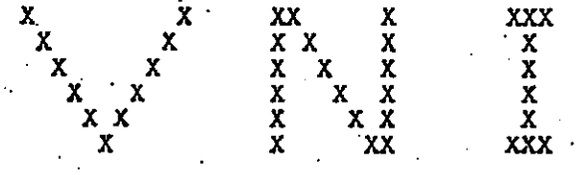
En dehors des variables, l'histoire de l'élaboration des fichiers de travail peut être aussi suffisamment longue et complexe pour susciter des éclaircissements par traitement de texte.

Il en va ainsi, pour la vague 1985, de la constitution du fichier de travail des ménages, qui a transité par 17 stades.

Un sous-document retraçant l'historique de l'élaboration de ce fichier est présenté en illustration (page 52).

2.2.1 - Document 1 : V.N.I.(Variables Nouvelles Individuelles)

V.N.I. page 1



VARIABLES NOUVELLES INDIVIDUELLES

Avertissement

Par souci de confidentialité, tous les documents de travail présentés à titre illustratif ont été épurés des noms des fichiers et des programmes de création (remplacés par des XXXXXXX).

PRESENTATION

Ce document se propose de présenter les VARIABLES NOUVELLES du FICHIER INDIVIDUEL panel ménage 1985.
Il complète utilement la base de données VARIN85B qui répertorie la totalité des variables de ce fichier, originales et créées, et est à consulter conjointement.

Dans la base de données on pourra trouver, pour les 620 variables d'origine comme pour les 190 variables créées à partir de celles-ci, des spécifications relatives à leur format, à leur contenu (ventilé sous 34 rubriques), à leur type (variable d'origine, nouvelle ou d'apport extérieur), au programme informatique qui les a créées et au fichier SPSS-X où elles apparaissent pour la première fois.
Toutes ces variables sont présentes et exploitables actuellement dans le fichier SPSS-X de nom XXXXXXXX qui constitue la dernière mise à jour du fichier individuel.

Le présent document se rapporte exclusivement aux 190 variables nouvelles et a pour fin de préciser l'histoire de leur création. Cette histoire étant susceptible de présenter de très grandes variations d'une variable à l'autre (allant par exemple d'un simple "recode" à la prise en compte de trois fichiers), il n'a pas semblé pertinent de suivre une systématisation absolue pour la présenter.

La présentation respecte toutefois les règles suivantes:

- Pour chacune des variables, on dispose de son nom (VARNAME), de son libellé (VARLABEL) et de son format (tel qu'il est déclaré dans la dernière mise à jour du fichier SPSS).

- L'ordre de présentation suit l'ordre des variables du fichier SPSS qui structure naturellement les variables par grands types (cf. PLAN page suivante).

PLAN

LES IDENTIFIANTS

NBMEN format F4 Numéro d'ordre du ménage

C'est la simple reprise de la V002 (formés par data list) du programme SPSS : FILE TYPE NESTED CASE = NBMEN.

CASEQ format F8 Numéro d'identification (n. d'ordre)

C'est le numéro d'ordre incrémenté de l'individu, qui permettra son suivi dans le panel.

IDEN format A13 Numéro composite d'identification individu

Il est formé par la concaténation de :
-l'année d'enquête : V001 à 2 caractères
-le n. du ménage : V002 = NBMEN à 4 caractères
-le n. du groupe : V504 à 2 caractères.
-le n. individuel : V505 à 2 caractères;
chacun de ces éléments étant séparé par un tiret.
On obtient par exemple, pour le premier individu du premier groupe du premier ménage, l'IDEN suivant :
"85- 1- 1- 1"

DESCRIPTION DEMOGRAPHIQUE

1-l'âge

DEM001 Format F8.2 Age en continu

C'est l'âge du sujet, en années, au 30 avril de l'année considérée (30-04-85). Il est calculé à partir de la fonction YRMODA de SPSS. Si on ne connaît pas le jour de naissance, on considère que le sujet est né le 15; si, exceptionnellement, on ignore le mois de naissance, on attribue le mois de juin. Le programme SPSS est le suivant:

```
COMPUTE YV510=V510 (année de naissance)
COMPUTE YV511=V511 (mois de naissance)
IF V511 EQ -9 YV511=6
IF V510 EQ -9 YV510=15
COMPUTE DATEREF=YRMODA(1985,4,30)
COMPUTE AG=(DATEREF-YRMODA (YV510,YV511)/365,25
COMPUTE VIC001=TRUNC(AG)
```

On ne conserve que la VIC001 qui est renommée DEM001

DEM002 Format F8.2 Age simplifié

Regroupement des jeunes de moins de 15 ans par RECODE de l'ancienne VIC001 = DEM001

```
   labels 1  moins de 15 ans
          15  15 ans
          16  16 ans
          etc...
```

DEM003 Format F8.2 Classes d'âge quinquennales complètes

Regroupement de l'âge en continu en 20 classes (par RECODE de la VIC001)

```
   labels  1  0 à 4 ans
           2  5 à 9 ans

          20  95 ans et plus
```

Ancienne VIC003 renommée.

DEM004 Format F8.2 Classes d'âge quinquennales recodées (début)

Regroupement des 0 à 14 ans en une seule catégorie (par RECODE de la VIC003).

```
   labels  1  0 à 14 ans.
           2  15 à 19 ans

          18  95 ans et plus
```

Ancienne VIC004 renommée.

DEM005 Format F8.2 Grands groupes d'âge type 1

Formation de trois catégories (15-60) par RECODE de la VIC001=DEM001
labels 1 de 0 à 14 ans
2 de 15 à 59 ans
3 60 ans et plus
Ancienne VIC005 renommée.

DEM006 Format F8.2 Grands groupes d'âge type 2

Formation de trois catégories (20-65) par RECODE de la VIC001=DEM001
labels 1 de 0 à 19 ans
2 de 20 à 64 ans
3 65 ans et plus
Ancienne VIC006 renommée.

DEM007 Format F8.2 Enfance,âge d'activité,retraite

Formation de trois catégories (15-60) par RECODE de la VIC001=DEM001
labels 1 de 0 à 15 ans-enfance
2 de 16 à 64 ans-age d'activité
3 65 ans et plus
Ancienne VIC007 renommée.

DEM008 Format F8.2 Age référence population nubile

Formation différentielle selon le sexe par RECODE de la VIC001=DEM001, selon la V509
labels 0 non concerné
1 homme de 18 ans et plus
2 femme de 16 ans et plus
Ancienne VIC008 renommée.

DEM009 Format F8.2 Age C.S.B.

RECODE de la VIC001=DEM001 labels 0 de 0 à 15 ans
1 de 16 à 24 ans
Ancienne VIC009 renommée. 2 de 25 à 49 ans
3 de 50 à 64 ans
4 de 65 à 74 ans
5 75 ans et plus

DEM010 Format F8.2 Age C.S.B. : actif-non actif

Regroupement différentiel selon le sexe.Sont actifs :
-les hommes de 25 à 64 ans
-les femmes de 25 à 59 ans
-les jeunes de moins de 25 ans qui ont un emploi ou perçoivent le chômage ou ont déjà travaillé.
labels 1 enfants
2 actifs
3 personnes âgées
Le programme prend en compte l'âge en continu (VIC001), le sexe (V509), l'âge d'entrée dans la vie professionnelle (V883), le fait d'avoir un emploi en avril (V805) et le fait de percevoir le chomage (V854).
Ancienne VIC010 renommée.

DEM011 Format F8.2 Classes d'âge quinquennales recodées (fin)

Regroupement des 3 catégories d'âge à partir de 85 ans, par RECODE de l'ancienne VIC003
labels 1 de 0 à 4 ans
2 de 5 à 9 ans

18 85 ans et plus
Ancienne RVIC003 renommée.

2 - la structure familiale

DEM012 Format F8.2 Situation familiale

Regroupement des catégories "séparés de fait" et "séparés de droit" par RECODE de la V512.
Ancienne VIC030 renommée.

DEM013 Format F8.2 Enfant-adulte

Différenciation selon l'âge légal de la majorité (18 ans) par RECODE de l'ancienne VIC001 = DEM001
Labels 1 enfant : moins de 18 ans
2 adulte : 18 ans et plus
Ancienne VIC011 renommée.
Cette variable est à la base de la création des deux indicateurs qui suivent.

DEM014 Format f8.2 Indicateur : enfant de moins de 18 ans

Créé d'après la DEM013
Ancienne VIC012 renommée

DEM015 format F8.2 Indicateur adulte : plus de 18 ans

créé d'après la DEM013
ancienne VIC013 renommée

DEM016 format F8.2 Indicateur : partie d'un couple (marié ou non)

Un couple est défini comme 2 adultes, chacun d'eux ayant un conjoint dans le ménage. On est membre d'un couple si le n. du conjoint (ou ami) est déclaré (V597) dans le questionnaire C complet.
Ancienne VIC014 renommée.

DEM017 format F8.2 Indicateur : questionnaire C simplifié

Indicateur créé sur la base de la V532 après vérification.
Ancienne VIC015 renommée

DEM018 format F8.2 Indicateur : questionnaire C complet

Indicateur créé sur la base de la V532 après vérification.
Ancienne VIC016 renommée

DEM019 format F8.2 Indicateur : partie d'un couple marié

De par la définition restrictive du couple, il s'agit exclusivement d'adultes.
Conditions à remplir :
- être marié (d'après la V512 du tableau familial)
- avoir un numéro de conjoint ou ami..
Ancienne VIC031 renommée.

DEM020 format F8.2 Indicateur : apparenté au chef de ménage

Il s'agit de toute la famille légale (sang ou alliance). Pour une liste des cas concernés, on se reportera au tableau présenté page suivante.
Ancienne VIC032 renommée.

DEM021 format F8.2 Indicateur : non apparenté au chef de ménage

Selon le "lien avec le chef de ménage", les conditions à remplir sont strictement exclusives de celles qui conduisent à la DEM020 (cf. tableau page suivante). L'indicateur concerne en particulier la compagne (épouse non légale) et sa famille, l'enfant à charge parce que placé dans ce ménage.

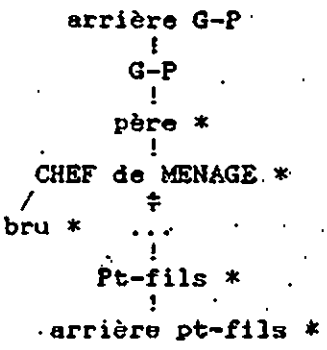
LIEN AVEC LE CHEF DE MENAGE

Valeur dans le tableau de composition du ménage		DEM020 apparenté	DEM021 non apparenté	TROISG trois générations	DEM029 ascendant
CHEF de MENAGE	1				
épouse	2	X			
"épouse"	3		X		
ami(e)	4		X		
fils (fille)	5	X			
beau-fils (fille à	6	X			
"beau-fils"	7		X		
bru (gendre)	8	X		X	
fils adoptif	9		X		
frère (soeur)	10	X			
beau-frère	11	X			
"beau-frère"	12		X		
père (mère)	13	X		X	X
beau-père	14	X		X	X
"beau-père"	15		X		
pt-fils (fille)	16	X			
arrière pt-fils	17	X			
gd-père (mère)	18	X		X	X
arrière gd-père	19	X		X	X
gd père de l'épse	20	X		X	X
arr.gd-père épse	21	X		X	X
neveu (nièce)	22	X			
neveu de l'épouse	23	X			
oncle (tante)	24	X			
oncle de l'épouse	25	X			
cousin(e)	26	X			
cousin(e) de l'épse	27	X			
fils(le) de l'ami	28		X		
mari	29	X			
relation C.M. *	30	X			
relation épouse *	31	X			
relation compagne	32		X		
non relation C.M. **	33		X		

(*) Toute personne apparentée au C.M. ou a l'épouse ne figurant pas dans le tableau sous les n. 1 à 29. Par exemple : gd-tante.
(**) Toute personne non apparentée au C.M. vivant toutefois sous son toit. Par exemple : couple homosexuel.

DEM022 format F8.2 Indicateur : partie d'une lignée de 3 générations

Indique que la personne fait partie d'une lignée ascendante directe de 3 générations.
Exemple : les individus marqués d'une (*) dans le schéma ci-contre sont susceptibles de prendre la valeur 1, ayant 2 générations au moins au dessus d'eux.
La présence de la bru ou du gendre s'explique pour témoigner d'une ligne descendante (de substitution) en raison de l'interruption.



TROISG format F8.2 Indicateur : 3 générations - restraint

Cet indicateur reprend la DEM022 en excluant les cas où la personne remplit les conditions suivantes : être marié; avoir un conjoint ou ami dans le ménage, être apparenté au chef de ménage (C.M.). De fait, il s'agit principalement d'exclure l'épouse du C.M.
L'indicateur est construit à partir des variables V512,V597 et DEM020.
Cette variable sert la construction d'une typologie familiale spécifique adoptée lors de l'exploitation du fichier 1985.Elle ne sert que cet usage particulier.

DEM023 format F8.2 Indicateur : a père (ou mère) dans le ménage

Concerne toute personne des questionnaires C complet ou simplifié ayant au moins l'un de ses deux parents dans le ménage. Est créé à partir des variables V531,V532,V595,V596.
Ancienne VIC035 renommée.

DEM024 format F8.2 Indicateur : a père (mère) et est apparenté au C.M

C'est une personne apparentée au chef de ménage dans le sens de la DEM023 et qui a son père ou sa mère dans le ménage, au sens de la DEM020.
Ancienne VIC036 renommée.

DEM025 format F8.2 Indicateur : épouse du chef de ménage.

Indicateur créé à partir de la V507 (lien avec le chef de ménage).
Ancienne VIC037 renommée

DEM026 format F8.2 Compteur déductif pour ménage

Cet indicateur est positif si l'un des deux indicateurs précédents (DEM024, DEM025) est positif. Il signale donc que la personne :
- soit est l'épouse du C.M.
- soit est apparentée au C.M. et a l'un de ses parents au moins dans la ménage.

Cet indicateur sert principalement l'établissement des typologies de ménage.

DEM027 format F8.2 Indicateur : célibataire, a père (mère)

Signale que la personne est célibataire (d'après la V512) et qu'elle a son père et/ou sa mère dans le ménage (selon la DEM023).
Ancienne VIC039 renommée.

DEM028 format F8.2 Indicateur : non célibataire, a père (mère)

La personne est non célibataire (mariée, veuve, divorcée ou séparée selon la V512) et elle a son père et/ou sa mère dans le ménage (selon la DEM023).
Ancienne VIC040 renommée.

DEM029 format F8.2 Indicateur : ascendant du C.M. (R.M.G.

La définition de l'ascendant est celle qui est prise en compte dans l'établissement du revenu minimum garanti. Est concerné tout ascendant - y compris ascendant de la compagne ou amie - qui remplit la condition d'être une personne âgée (non active : DEM010 = 3).

DEM030 format F8.2 Indicateur : impotent (R.M.G.)

Cet indicateur correspond à la définition d'une personne impotente dans le sens de l'établissement du revenu minimum garanti.
Est impotente toute personne qui :
- soit a un handicap quel qu'il soit (V529)
- soit est invalide à 50 % au moins (V530)

MATCJ format F8.2 Matricule du conjoint

Chaque membre d'un couple est assorti du numéro individuel (CASEQ) de son conjoint. Cette variable est créée selon la logique générale du report d'information inter-membres d'un ménage (cf.annexe).
Cette création fait intervenir conditionnellement les variables DEM016,V505,V597. Elle nécessite le passage par un fichier agrégé au niveau du ménage, suivi d'un déversement des informations sur les cas individuels concernés.

MATENF1 format F8.2 Matricule de l' enfant n. 1
: :
MATENF10 format F8.2 Matricule de l'enfant n. 10

Les variables MATENF1 à MATENF10 contiennent, au niveau des enregistrements de la mère et/ou du père, le numéro CASEQ de chacun de ses enfants, indépendamment du fait que l'enfant soit à charge ou non. C'est le lien de filiation qui est cerné.
Cette création fait intervenir conditionnellement les variables DEM017,DEM018,V531,V532,,V595,V596. Elle nécessite le passage par un fichier agrégé au niveau du ménage, suivi d'un déversement des informations sur les cas individuels concernés.

DEM031 format F8.2 Nombre d'enfants dans le ménage

Il s'agit du simple dénombrement des enfants dont le CASEQ est enregistré au niveau du père et de la mère par les variables MATENF1 à MATENF10. "Enfant" est pris dans le sens général de lien de filiation.

CSPL format F8.2 Repère sujet C simplifié ayant un père

La variable est créée comme suit :
- le questionnaire C simplifié fait mention d'un père
- et le questionnaire C complet ne fait pas mention d'un père.

Ce sujet, qui remplit un questionnaire C simplifié a donc un père dans le ménage.

CCPL format F8.2 Repère sujet C complet ayant un père

Concerne les enregistrements où un numéro de père est indiqué si le sujet est un C complet et ne l'est pas si le sujet est un C simplifié.

SEQP format F8.2 Numéro du père dans le ménage

Dans le cas où le sujet a son père et/ou sa mère dans le ménage (selon la DEM023), le numéro du père dans le tableau familial (V531 ou V595) est attribué, selon que le sujet est "C simple" ou "C complet" (d'après CSPL ou CCPL).

SEQM format F8.2 Numéro de la mère dans le ménage

dans le cas où le sujet a son père et/ou sa mère dans le ménage (selon la DEM023), le numéro de la mère dans le tableau familial (V532 ou V596) est attribué, selon que le sujet est "C simple" ou "C complet".
Même logique de création que la SEQP.

SEQINC format F8.2 Repère des non célibataires ayant père et/ou mère dansle ménage

Si le sujet est non célibataire (selon DEM028) et qu'il a son père et/ou sa mère dans le ménage, on le repère par son numéro individuel (V505). Cette variable est la première étape d'un report d'information inter-membres d'un ménage.

SCOLARITE ET POSITION PROFESSIONNELLE

1. SCOLARITE

SCOL1 format F8.2 Indicateur scolaire fichier S

Concerne les C simplifiés.
Regroupement, par "RECODE" de la variable V559 : type d'enseignement. Les nouvellles valeurs sont :

0	non scolarisé	5	les 3 filières techniques
1	jardin d'enfant	6	enseign.complémentaire
2	primaire ordinaire	7	les 2 enseign.supérieurs
3	toutes classes spéc.	8	apprenti en technique
4	enseign.secondaire		

La valeur 8 est attribuée aux jeunes qui suivent un enseignement technique et perçoivent en avril un salaire d'apprenti (V582).

SCOL2 format F8.2 Indicateur scolaire fichier C + apprenti

Concerne les "C complet", en particulier les adultes en formation.
Regroupement par "RECODE" de la V652 : type d'enseignement adulte.Les nouvelles valeurs sont les suivantes :

5	enseign.secondaire et 3 filières techniques
7	deux types d'ensegnement supérieur
8	apprentissage adulte (si le sujet perçoit un salaire d'apprentissage selon la V866).

SCOL3 format F8.2 Indicateur de formation scolaire achevée

Concerne les "C complet"
Regroupement par "RECODE" de la V681 : formation scolaire et professionnelle pour un C complet. Nouvelles valeurs :

0	pas de formation	3	formation secondaire
1	primaire ou complémentaire	4	formation supérieure
2	technique ou professionnelle		

NB: corriger le label de la valeur -5 : C simplifié

SCOL4 format F8.2 Indicateur de formation achevée ou en cours

- S'il s'agit d'un C simplifié (SCOL1 différent de -5), on reprend la formation en cours (SCOL1) que l'on recode en catégories moins fines.
- S'il s'agit d'un adulte en formation (SCOL2 > 0), on considère sa formation achevée (SCOL3) et l'on recode :
  - les "primaire" en "professionnel et technique"
  - les "non réponse" en "professionnel et technique" (2 cas seulement sont concernés).
- Si l'adulte n'est pas en formation; on reprend simplement la formation achevée : SCOL3.

Les valeurs sont :

-9 N.S.P.		2 techn.profes.
0 pas de formation		3 secondaire
1 primaire		4 supérieure

2. PROFESSION

Les variables JOB001 à JOB025 concernent le prestige professionnel et la catégorie socio-professionnelle.
Créées originellement sous le nom RV...,le programme de création les a sauvées directement en tant que JOB...

Les variables JOB001 à JOB015 concernent toutes des niveaux plus ou moins fins du "prestige socio-professionnel". Ce construit de prestige a été opérationnalisé par l'échelle de TREIMAN (Occupational Prestige in Comparative Perspective, New York, Academic Press, 1977),sur la base des différentes variables indicatrices des professions codées selon la classification internationale du type des professions.

Le code de PRESTIGE PROFESSIONNEL est établi à partir des variables suivantes :
- V657 profession exercée par un adulte en cours d'activité.
- V714 activité antérieurement exercée, pour un retraité.
- V884 premier emploi.
- V885 dernier emploi exercé.
- V887 profession du père quand l'adulte entrait dans la vie active.

Ces 5 variables de base ont permis la création de 3 groupes de variables JOB...;de plus, elles ont conduit à la formation de 2 groupes de variables JOB... concernant la catégorie professionnelle selon le tableau suivant :

	variables BIT prestige			variables catégorie prof.	
V657	JOB001	JOB006	JOB011	JOB016	JOB021
V714	JOB002	JOB007	JOB012	JOB017	JOB022
V884	JOB003	JOB008	JOB013	JOB018	JOB023
V885	JOB004	JOB009	JOB014	JOB019	JOB024
V887	JOB005	JOB010	JOB015	JOB020	JOB025

JOB001 à JOB005 format F2.0 Prestige B.I.T. ...

Le groupe de variables JOB001 à JOB005 correspond au "RECODE"des variables de base et conduit à la création des valeurs :

11 comptable

99 manoeuvre NCA

JOB006 à JOB010 format F2.0 Prestige B.I.T. recode1 ...

le groupe de variables JOB006 à JOB010 correspond au regroupement des valeurs des variables JOB001 à JOB005 en 7 catégories de 10 :

1 de 10 à 20
2 de 21 à 30
.
.
6 de 61 à 70
7 71 et plus

JOB011 à JOB015 format F2.0 Prestige B.I.T. recode2 ...

Le groupe de variables JOB011 à JOB015 correspond à un regroupement des valeurs des variables JOB006 à JOB010 en 3 grandes catégories :

1 de 10 à 35
2 de 36 à 45
3 46 et plus

JOB016 à JOB020 format F2.0 B.I.T.catégorie principale ...

Les variables JOB016 à JOB020 sont établies en ne retenant que la première valeur des variables de base : aux "blancs" est attribuée la valeur de donnée manquante (-9,-8 ou -5); les valeurs 0 et 1 sont recodées 1; les valeurs 7 à 9 sont recodées 7. On aboutit aux 7 catégories suivantes :

1 sciences,technique,prof.libérales
2 directeurs,cadres
3 administration
4 commerce, vente
5 services
6 agriculture
7 ouvriers

JOB021 à JOB025 format F2.0 B.I.T.catégorie secondaire ...

Les variables JOB021 à JOB025 correspondent aux deux premières valeurs des variables de base.

ANCRAGE

ANC01 format f8.2 Pays de naissance code A

Recode de la variable pays (V516) en 2 catégories:
1 Luxembourg
2 autre

ANC02 format f8.2 Pays de naissance code B

Recode de la variable pays (V516) en 3 catégories:
1 Luxembourg
2 C.E.E
3 autre.

ANC03 format f8.2 Pays de naissance code C

Recode de la variable pays (V516) en 3 catégories:
1 Luxembourg
2 limitrophe
3 autre

ANC04 format f8.2 Langue maternelle A

Recode de la variable langue (V517) en 2 catégories:
1 Luxembourg
2 autre

ANC05 format f8.2 Langue maternelle B

Recode de la variable langue (V517) en 6 catégories:
1 Luxembourg
2 allemand
3 français
4 italien
5 portugais
6 autre

ANC06 format f8.2 Nationalité A

Recode de la variable nationalité (V518) en 2 catégories:
1 Luxembourgeoise
2 autre

ANC07 format f8.2 Nationalité B

Recode de la variable nationalité (V518) en 3 catégories:
1 Luxembourgeoise
2 C.E.E
3 hors C.E.E

ANC08 format F8.2 Année de naturalisation

Recode de l'année de naturalisation (V519) par tranches de 10 ans :
1 jusqu'à 1930
2 de 1931 à 1940
.
.
6 de 1971 à 1980
7 1981 et au delà

Les 3 variables ANC09, ANC10 et ANC11 concernent la connaissance des 3 langues les plus usitées au Luxembourg (luxembourgeois, allemand, francais) et sont construites selon la même logique, à partir des deux variables concernant l'expression et la compréhension de chacune des langues.

ANC09 format F8.2 Connaissance du luxembourgeois

Calculée à partir des variables V521 et V522

ANC10 format F8.2 Connaissance de l'allemand

Calculée à partir des variables V523 et V524

ANC11 format F8.2 Connaissance du francais

Calculée à partir des variables V525 et V526

Le code 8 : N.A.P. des variables de base V521 à V526 est transformé exceptionellement en code 4 pour établir un ordonnancement croissant d'intervalle constant sur les valeurs de ces variables. Les valeurs des variables de base sont dès lors :
1 sans difficulté (à comprendre/à s'exprimer)
2 quelques difficultés.
3 beaucoup de difficultés
4 trop jeune ou handicap

La logique de construction est la suivante:
- Si aucune information n'est disponible, tant en expression qu'en compréhension, le code est 9 = N.S.P.
- Si une seule des deux informations est disponible, le calcul est effectué sur cette seule base.
- La connaissance de la langue, pour chacune des trois langues est calculée comme étant la moyenne "expression, compréhension", arrondie à l'unité supérieure.

Les codes finaux de ces variables "connaissance de la langue", sont ventilés dans le tableau croisé suivant:

compréhension \ expression	1	2	3	4
1	1	2	2	-
2	2	2	3	-
3	2	3	3	4
4	-	-	4	4

compréhension \ expression	2	4	6	8
2	1	2	3	-
4	2	3	4	-
6	3	4	5	6
8	-	-	6	7

On obtient ainsi une échelle de connaissance de 1 à 4 où 1 représente la meilleure connaissance et 4 la plus mauvaise.

Remarque : Il serait intéressant de multiplier les codes de base par 2 puis de soustraire 1 à la valeur finale, ce qui donnerait une échelle plus fine de 1 à 7 où les valeurs seraient susceptibles de prendre un "label" précis (cf.T.C de droite).

REVENUS

Les groupes de variables REV...,MDREV.... et AREV... se rapportent aux revenus.

- Les variables REV1 à REV25 sont les revenus individuels divers du mois d'avril 1985. Si leur perception n'est pas mensuelle, ils ont été calculés au prorata de leur périodicité de perception pour être ramenés à des revenus mensuels.

- Les variables MDREV1 à MDREV25 sont des indicateurs de revenus inconnus (lire "Missing Data" sur les REVenus). En cas de revenu inconnu, c'est à dire quand l'indicateur MDREV... vaut 1, le revenu le plus probable a été imputé sur la base du calcul de la médiane ou après calcul de régression multiple.

- Les variables AREV1 à AREV25 correspondent à des indicateurs de revenu non nul. Cet indicateur prend la valeur 1 quand le revenu correspondant est non nul.

- Quant aux variables REV26 à REV42, elles renvoient à différents totaux effectués sur les variables REV... de base.

Suit un tableau récapitulatif des variables REV1 à REV25, et des variables MDREV... et AREV... qui leur correspondent. Ce tableau présente:
-le libellé du type de revenu
-la variable originale qui lui sert de support.
-le type de questionnaire auquel il se rapporte (C pour complet, S pour simplifié)
-le nom de la variable REV...
-le nom de la variable MDREV...
-le nom de la variable AREV...
-le type d'imputation qui a été effectué pour les données manquantes.

Ce tableau contient les informations élémentaires; il est suivi de l'explicitation des quelques éléments qui peuvent poser problème.

V.N.I. page 22

VARIABLES DE REVENUS

Salaire mensuel net	V729	C REV1	MDREV1	AREV1	régression
Primes de fin d'année	V735	C REV2	MDREV2	AREV2	régression
Salaires en nature	V741	C REV3	MDREV3	AREV3	Méd: 850
Activité prof.accessoire	V747	C REV4	MDREV4	AREV4	Méd: 10800
Bénéf.exploit.agricole	V753	C REV5	MDREV5	AREV5	individu
Bénéf.indust.commerce	V765	C REV6	MDREV6	AREV6	Méd: 50000
Bénéf.non commerciaux	V771	C REV7	MDREV7	AREV7	Méd: 70000
Pension vieillesse	V777	C REV8	MDREV8	AREV8	régression
Pension de survie	V783	C REV9	MDREV9	AREV9	régression
Pension alimentaire	V789	C REV10	MDREV10	AREV10	NAP en 85
Pension d'invalidité	V801	C REV11	MDREV11	AREV11	Méd: 24700
Prestations spéciales	V808	C REV12	MDREV12	AREV12	NAP en 85
Prestation supplémentaire	V813	C REV13	MDREV13	AREV13	Méd: 6000
Fond National Solidarité	V819	C REV14	MDREV14	AREV14	Méd: 9000
Rente viagère	V825	C REV15	MDREV15	AREV15	Méd: 12000
Rente d'associat.privée	V831	C REV16	MDREV16	AREV16	NAP en 85
Rente accident (perman)	V843	C REV17	MDREV17	AREV17	Méd: 23500
Pension d'orphelin	V578	S REV18	MDREV18	AREV18	Méd: 7000
Prest.incapacité travail	V849	C REV19	MDREV19	AREV19	NAP en 85
Indemnité chômage	V855	C REV20	MDREV20	AREV20	Méd: 21500
Bourses d'étude	V861 V594	C REV21 S	MDREV21	AREV21	NAP en 85
Salaire d'apprentissage	V867 V589	C REV22 S	MDREV22	AREV22	Méd: 10000
Solde du milicien	V879	C REV23	MDREV23	AREV23	NAP en 85
Salaire d'un job enfant	V573	S REV24	MDREV24	AREV24	Méd: 4500

TOTAL ou NOMBRE ("sum" ou "count") REV25 MDREV25 AREV25

PERIODICITE DE PERCEPTION DES REVENUS

Tous les revenus ont été ramenés à un revenu mensuel en prenant en compte la périodicité de leur perception, variable qui suit immédiatement celle qui est présentée dans le tableau précédent (qui indique le montant du revenu correspondant).

Ainsi, si le revenu déclaré est mensuel, il est pris dans sa totalité.

S'il est trimestriel le montant est divisé par 3.
S'il est semestriel son montant est divisé par 6.
S'il est annuel il est divisé par 12.

Quelques cas sont litigieux:

REV18 Pension d'orphelin
Quel que soit le mois de sa perception, le montant a été pris dans sa totalité comme un revenu du mois d'avril.
Si la périodicité était exceptionnelle (code 7) et que sa perception tombait en dehors de la période de référence (janvier à avril) (code -9 pour son montant), le montant a été porté à 1 franc (franc symbolique).

REV21 Bourse d'étude:
Si elle est perçue en dehors des 4 mois de référence (janvier à avril), on considère qu'elle est annuelle et on divise son montant par 12.
Si elle est perçue 4 fois, elle est mensuelle.
Si elle est perçue 1 fois sur la période des 4 mois de référence (on l'a divisée par 12) mais elle pourrait aussi être trimestrielle...

REV22 Salaire d'apprentissage:
Dans le cas du salaire d'apprentissage perçu par un individu "C simplifié", on ne connaissait pas la périodicité. Dans tous les cas, quel que soit le mois de perception (janvier à avril), le montant a été considéré comme salaire d'avril.

REV24 Job d'un enfant:
Quel que soit le mois de sa perception , le montant a été considéré dans son entier comme salaire du mois d'avril.

Remarque: Le programme de création de ces variables ne prévoit la valeur 7 (exceptionnelle) comme périodicité de perception que si un cas au moins s'appliquait en 1985.
Lors de sa réutilisation dans les années ultérieures, il faudra penser à vérifier systématiquement la présence de ce code.

TOTAUX DIVERS SUR LES REVENUS

Sur la base des variables REV1 à REV24, diverses sommes de revenus sont calculées. Elles sont présentées dans le tableau qui suit.

variables SOMMES de revenus

	REV25	REV26	REV27	REV28	REV29	REV30	REV31	REV32	REV33	REV34	REV35	REV36	REV37	REV38	REV39	REV40 *	REV41 *	REV42 *
REV1	X	X	X		X	X									X	X	X	X
REV2	X	X	X		X	X									X	X	X	X
REV3	X	X	X		X	X									X	X	X	X
REV4	X		X		X	X									X	X	X	X
REV5	X			X	X	X									X	X	X	X
REV6	X			X	X	X									X	X	X	X
REV7	X			X	X	X									X	X	X	X
REV8	X						X	X	X				X	X	X	X	X	X
REV9	X						X	X	X				X	X	X	X	X	X
REV10	X						X	X	X				X	X	X	X	X	X
REV11	X						X	X	X				X	X	X	X	X	X
REV12	X						X	X	X				X	X	X	X	X	X
REV13	X						X	X	X				X	X	X	X	X	X
REV14	X						X	X					X				X	
REV15	X						X								X	X	X	X
REV16	X						X								X	X	X	X
REV17	X						X	X	X				X	X	X	X	X	X
REV18	X						X	X	X				X	X	X	X	X	X
REV19	X									X	X	X	X	X	X	X	X	X
REV20	X									X	X	X	X	X	X	X	X	X
REV21	X									X	X				X	X	X	X
REV22	X					X				X					X	X	X	X
REV23	X									X	X				X	X	X	X
REV24	X					X									X	X	X	X

Nota: Les variables REV40, REV41, REV42 sont spécifiques aux ascendants: Elles ont une valeur si la DEM029 vaut 1. Dans ce cas:
- REV40 correspond à REV39
- REV41 correspond à REV25
- REV42 correspond à: REV40 - 12361 F. (R.M.G.)

Les libellés de ces variables-sommes, ainsi que leur format sont les suivants.

NOM	FORMAT	LIBELLE
REV25	F8.2	Revenu individuel total (des 24)
REV26	F8.2	Total revenu salarial activité principale
REV27	F8.2	Total revenu salaires: act.principale & accessoire
REV28	F8.2	Total bénéf.princip.:agricole,industriel,indépdt
REV29	F8.2	Total revenu travail: sal.& bénéf. princip.& access.
REV30	F8.2	Total revenus travail & apprentissage & job
REV31	F8.2	Total des pensions, retraites etc...
REV32	F8.2	Total des pensions, retraites par securité sociale
REV33	F8.2	Total des pensions, retraites par sec.soc hors FNS
REV34	F8.2	Total rev. remplacement transit sauf job
REV35	F8.2	Total rev. remplacement hors apprentissage
REV36	F8.2	Total rev.remplacement transit sec soc
REV37	F8.2	Total rev transferts permanent ou accessoire
REV38	F8.2	Total rev.transferts permanent ou access. hors FNS
REV39	F8.2	Revenu total hors FNS
REV40	F8.2	Revenu immunisé FNS, pour ascendant
REV41	F8.2	Revenu non immunisé FNS, pour ascendant
REV42	F8.2	Revenu immunisé FNS, pour ascendant, moins R.M.G.

ENRICHISSEMENTS

1- RECHERCHE "MIR"

Les variables SELMIR1, SELMIR2, IDMIR, ENFMER, TOTEAMIR et CRITCO sont quelques variables qui ont été séléctionnées dans un fichier créé à partir du fichier panel et servant la recherche "MIR" pour venir enrichir le fichier individuel.

Elles sont considérées comme un apport extérieur, la recherche "MIR" poursuivant un objectif spécifique; le détail de leur création n'est pas rapporté ici. Pour une documentation complète on pourra se reporter au programme de création de XXXXXX (fichier SPSS-X).

SELMIR1 format F1 Critère général de sélection MIR

Ce critère définit les sujets qui sont potentiellement intéressants dans le cadre de l'étude MIR, à savoir les femmes de 17 à 54 ans.
Créé à partir des variables DEM001 et V509.

SELMIR2 format F1 Critère final de sélection MIR

Critère de sélection des femmes de 17 à 54 ans ayant un enfant à charge dans le sens du projet MIR.
La création de la variable suit la logique du report d'information inter-membres d'un ménage. Elle passe par la création de la variable ENFTMER et suit les étapes suivantes:
- repère des enfants à charge (ENFCHA)
- sélection de ceux qui ont une mère : ENFTMER
  identification par reprise du n. de la mère dans le tableau familial (variables MIRX1 à MIRX7)
- agrégation au niveau du ménage; les variables agrégées sont AMIR1 à AMIR7
- répercussion de cette information au niveau individuel.
- Si AMIR a une valeur, IDMIR=1
  - Si c'est une femme MIR et que IDMIR vaut 1, le critère final de selection est retenu et prend la valeur 1. Il vaut 0 dans tous les autres cas.

IDMIR format F1 Identifie mère avec enfant à charge-MIR

ENFTMER format F1 Repère enfant-mère MIR

Enfant à charge ayant sa mère dans le ménage.
Créé à partir des variables DEM016, DEM017, DEM018, REV25, V532 et V596.
Dans le sens "MIR", les enfants à charge comprennent les adultes gagnant moins de 20601 francs. Un "enfant à charge" ayant sa mère dans le ménage est retenu s'il ne fait pas partie d'un couple.

TOTEAMIR format F1 Total d'enfants définition MIR

Nombre total d' "enfants à charge"(ENFTMER) qu'a la mère MIR.
Créé par la fonction COUNT

CRITCO format Critère de cohabitation MIR

Repère les femmes de 17 à 54 ans faisant partie d'un couple, qu'elles soient mariées ou non.
Critère créé à partir des variables DEM016 et DEM019.

2- RECHERCHE SUR L'OFFRE DE TRAVAIL

Cette étude qui se propose la description multivariée du marché du travail au Grand Duché et la recherche des déterminants de l'offre de travail permet l'enrichissement du fichier individuel par des variables concernant les enfants à charge.

AGEEC1 format F8.2 Age de l'enfant à charge n.1

AGEEC10 format F8.2 Age de l'enfant à charge n.10

Ces 10 variables sont créées selon la logique des transferts d'information inter-individus, intra-ménage (cf.annexe1).
Un enfant à charge est un enfant du questionnaire C simplifié.
L'âge (DEM001)de chacun des enfants est rapporté au niveau de l'enregistrement du père d'une part, de la mère d'autre part.

On trouvera les programmes de création sous les désignations XXXXX, XXXXXXXX et XXXXXXX; et les "hot listings" dans le classeur "XXX XXXXXXXX".

ANNEXE 1

REPORT D'INFORMATION INTER-INDIVIDUS INTRA-MENAGE

LOGIQUE DU REPORT

1- Fichier individuel
Sous les conditions du problème traité, on crée un REPERE qui associe deux membres d'un ménage.
Par exemple, le REPERE enfant-mère, associé à chaque enregistrement "enfant", qui signifie "être enfant du n.x dans le tableau familial",
ou le REPERE conjoint-conjoint, associé à chaque conjoint et signifiant "être conjoint du n.x dans le tableau familial".
Dans tous les cas créer un REPERE revient à repérer un membre du ménage A au moyen d'une variable B.

2- On détermine la valeur maximale de la variable B. On sait par exemple que le n. dans le tableau familial de "mère de..." va jusqu'à 7.
On sait par ailleurs qu'en 1985 il y a 11 individus au plus dans un ménage. Pour l'année 85, la valeur maximale de A est 11.

3- On obtient de ce fait une matrice croisant A et B qui va permettre d'initialiser autant de variables que d'éléments dans cette matrice (à l'exception des éléments de la diagonale).

Par exemple on a croisé les 11 possibilités de A avec les 7 possibilités de B:

B:	1	2	3	4	5	6	7
A:							
1	X						
2	E12	X					
3	E13	E23	X				
4	E14	E24	E34	X			
5	E15	E25	E35	E45	X		
6	E16	E26	E36	E46	E56	X	
7	E17	E27	E37	E47	E57	E67	X
8	E18	E28	E38	E48	E58	E68	E78
9	E19	E29	E39	E49	E59	E69	E79
10	E110	E210	E310	E410	E510	E610	E710
11	E111	E211	E311	E411	E511	E611	E711

4- L'information d'intérêt est placée dans chacune des variables initialisées sous E12 à E711 comme suit:

```
Si B vaut 1
  si A vaut 2  alors E12=INFO
  si A vaut 3  alors E13=INFO
       .              .
       .              .
  si A vaut 11 alors E111=INFO

Si B vaut 2
  si A vaut 1  alors E21=INFO
  si A vaut 3  alors E23=INFO
       .              .
       .              .
  si A vaut 11 alors E211=INFO


Si B vaut 7
  si A vaut 1  alors E71=INFO
  si A vaut 2  alors E72=INFO
       .              .
       .              .
  si A vaut 11 alors E711=INFO
```

Sur les 7 enregistrements d'un ménage qui peuvent contenir une INFO, cette information placée dans les variables E12 à E711 est structurée comme suit:

```
enr1 xxxx
enr2    xxxx
enr3       xxxx          0
enr4          xxxx
enr5     0       xxxx
enr6                xxxx
enr7                     xxxx
```

5- Cette série d'information est concentrée sur un seul enregistrement, au niveau du ménage, par agrégation.
Pour chaque ménage est ainsi créé un enregistrement de type:

```
!-----------!-----------! . . . . . . . . . !-----------!
 E12     E111 E21     E211                   E71     E711
```

Seulement quelques valeurs sont présentes dans cette série de variables.

6- Sous les conditions de création des variables, l'information est déversée au niveau de l'individu:

Si A vaut 1, descente de l'INFO contenue dans E12 à E111 dans l'une des 10 variables définitives VARINFO1 à VARINFO10.

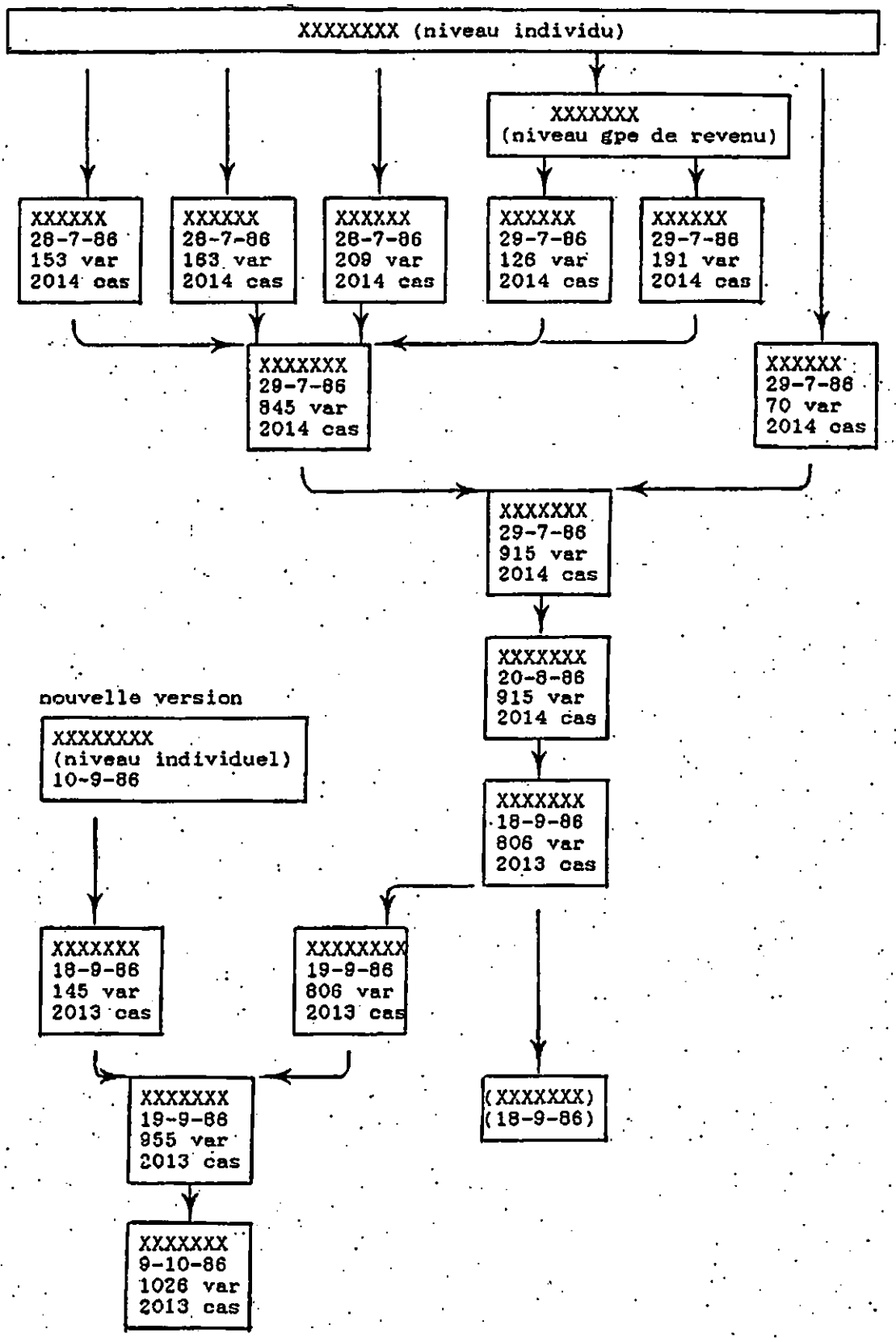
Si A vaut 2, descente de l'INFO contenue dans E21 à E211 dans l'une des 10 variables définitives VARINFO1 à VARINFO10.

etc...

Si A vaut 7, descente de l'INFO contenue dans E71 à E711 dans l'une des 10 variables définitives VARINFO1 à VARINFO10.

Au total, ce report d'informations inter-membres d'un ménage fait intervenir 3 fichiers.
- Le fichier individuel initial sert à repérer les cas concernés et l'information pertinente.
- Ces variables sont ensuite agrégées au niveau du ménage, constituant un fichier ménage.
- Puis l'information du niveau ménage est réperoutée au niveau individuel par la procédure MATCH FILES .

2.2.2 - Document 2 : Elaboration du fichier de travail sur les ménages.



2.3 - Documentation transversale par base de données.

Pour la vague 1985 ont été créées TROIS bases de données correspondant aux TROIS niveaux d'analyse (individu, groupe, ménage) et se rapportant aux TROIS fichiers de travail.

Ces bases de données, qui renseignent essentiellement sur les variables, sont de structure très similaire et pourraient être considérées comme UNE seule base, si les modalités actuelles du fonctionnement par rapport à trois fichiers devaient changer.

La troisième de ces bases n'étant pas terminée, les documents qui sont présentés dans ce chapitre ne constituent que des indications.

L'organisation finale en une ou plusieurs bases est affaire de fonctionnement.

Nous n'entendons pas faire ici autre chose qu'indiquer l'information contenue dans ces bases, indépendamment de leur organisation.

Leur vocation première est de renseigner sur les VARIABLES, non seulement sur leur contenu mais aussi sur leur localisation.

TOUTES les variables sont répertoriées dans ces bases qui contiennent trois groupes d'informations :

- l'identification et les caractéristiques de la variable, comme, par exemple, son nom, sa signification, son format, ...

- son contenu, c'est à dire son domaine d'appartenance (famille, revenu, événement de vie, logement, ...).

- des éléments de liaison avec les différents stades de son élaboration.

Sur ce dernier point, précisons que les variables, soit proviennent du QUESTIONNAIRE, soit sont créées par PROGRAMME. Dans tous les cas, elles appartiennent à des FICHIERS et elles sont renseignées dans divers DOCUMENTS écrits (bible de codification des variables, listage des programmes de création,...).

2.3.1 - Exemple illustratif : la base de données des variables individuelles.

La liste des informations contenues dans la base de données des variables du niveau individuel est présentée dans le tableau qui suit.

n°	NOM du CHAMP	CONTENU DU CHAMP
1	VARNAME	Nom de la variable dans le fichier.
2	FICHIND *	La variable est ou n'est pas une variable du fichier individuel.
3	FICHGPE	La variable est-elle reconduite au niveau du groupe de revenu ?
4	FICHMEN	La variable est-elle reconduite au niveau du ménage ?
5	VARLABEL	Signification de la variable dans le fichier SPSS.
6	QUESTION	Numéro de la question dans le questionnaire 1985.
7	NIVRECUEIL	Niveau du recueil (individu en général, enfant, adulte, groupe, ménage).
8	FORMAT	Format dans le fichier SPSS.
9	CODES	Renseigne sur les "value labels".
10	SUBOBJ	La variable est-elle de nature objective ou subjective.
11	TYPE	Est-ce une variable d'origine ou créée par programme ?
12	FICHIER	Fichier SPSS dans lequel la variable est utilisable.
13	PROGRAMME	Nom du programme qui l'a créée.
14	CONSULTER	Documents à consulter
15	REMARQUE	champ ouvert à toute remarque...
16	AIDE	contenu
17	ANCRAGE	contenu
18	BIENS_IMMO	contenu : biens immobiliers
19	CHANGEMENT	contenu
20	DEMOGRAPHI	contenu
21	DEPENSES	contenu
22	DIFFICULTE	contenu
23	EMPLOI	contenu
24	EMPRUNT	contenu
25	ENQUETE	contenu
26	EQUIPEMENT	contenu
27	EVENEMENT	contenu
28	FAMILLE	contenu
29	FINANCES	contenu
30	FORMATION	contenu
31	GARDE	contenu
32	GESTION	contenu
33	GROUPE_REV	contenu : groupe de revenu
34	IDENTIF	contenu
35	LOGEMENT	contenu
36	MONTANT	contenu

.../...

(*) Il s'agit d'une constante dans le cas de cette base de données et ce champ est présent dans les deux autres bases. C'est à ce titre que les 3 bases peuvent, dans l'état actuel de leur agencement, n'en constituer qu'une seule.

(suite)

n°	NOM du CHAMP	CONTENU DU CHAMP
37	PROFESSION	contenu
38	RELATIONS	contenu
39	RESSOURCES	contenu : tout revenu (de n°39 à 45)
40	REVAUTRES	contenu : autre revenu
41	REVCAPITAL	contenu : revenu du capital
42	REVPENSION	contenu : revenu de pension,...
43	REVPRESFAM	contenu : revenu prestations familiales
44	REVPROFIND	contenu : revenu de prof.indépendante
45	REVREMPLAC	contenu : revenu de remplacement
46	REVTRAVAIL	contenu : revenu du travail
47	SANTE	contenu
48	SCOLARITE	contenu
49	VIE_SOCIAL	contenu
50	VOITURE	contenu

Tous les champs relatifs à l'analyse de contenu sont des champs de type logique (de largeur 1 : "oui" ou "non").

Pour les autres champs, la structure du fichier est la suivante.

n°	NOM de CHAMP	TYPE	LARGEUR
1	VARNAME	Caractère	8
2	FICHIND	Logique	1
3	FICHGPE	Logique	1
4	FICHMEN	Logique	1
5	VARLABEL	Caractère	40
6	QUESTION	Caractère	6
7	NIVRECUEIL	Caractère	1
8	FORMAT	Caractère	5
9	CODES *	Caractère	15
10	SUBOBJ	Caractère	1
11	TYPE	Caractère	1
12	FICHIER	Caractère	8
13	PROGRAMME	Caractère	8
14	CONSULTER *	Caractère	10
15	REMARQUE	Caractère	15
..	.....	.....	..

Quelques remarques.

- Les CODES ne reprennent pas la totalité des "value labels", ce qui serait à la fois inutile (ces renseignements sont disponibles dans un dictionnaire

généré par SPSS) et impraticable du fait de la grande diversité du nombre des valeurs de chaque variable.
Ce champ fournit une indication sur les codes des valeurs de manière à ce que le consultant de la base se fasse une idée de la nature de la variable;
par exemple,

"échelle 7 pts"
"dichotomique"
"99 code prof.
"en francs"

- Pour répertorier les documents à CONSULTER, on s'accorde sur quelques abréviations des noms, telles que "BICO" pour BIble de COdification,
"DICT" pour DICTionnaire des variables,
...
Pour une homogénéisation de ce processus, tous les documents nouvellement créés portent un titre (ou une référence) à la fois bref et facile à mémoriser.
Il en va ainsi des documents internes comme V.N.I., V.N.G. et V.N.M. présentés sous 2.2 ou des documents PSELL-1, PSELL-2, PSELL-3, ... qui désignent tous les rapports et documents au niveau de la diffusion.

2.3.2 - Diverses utilisations de la base de données.

Ces bases de données peuvent être consultées à des fins multiples par les membres du service logistique ou par les analystes.

1) Pour un **renseignement spécifique**, on éditera directement l'enregistrement concerné qui fournira la réponse.
Il s'agira par exemple de répondre à des questions telles que :

- où est la variable V312 ?
- est-ce que telle variable a été reportée dans le fichier groupe de revenu ?
- quel programme a créé la variable V312 ?
- l'histoire de la variable est-elle documentée ?
- ...

2) Souvent, l'analyste souhaitera **obtenir une liste des variables** concernant un domaine sur lequel il doit travailler.
Par exemple, on pourra lister les noms des variables relatives au logement et obtenir une liste comme celle qui est présentée ci-dessous.

VARIABLES RELATIVES AU LOGEMENT

QUESTION	NOM de VARIABLE	SIGNIFICATION DE LA VARIABLE
A1	V007	statut d'occupation du logement
A2	V008	année de construction du logement
A3	V009	nombre de pièces ds le logement
A4	V010	eau chaude
A5	V011	eau froide dans le logement
A6	V012	W.C.à l'intérieur-cabinet distinct
A6	V013	W.C.à l'intérieur-salle de bain
A6	V014	W.C.hors logement mais réservé
A6	V015	W.C.hors logement, en commun
A6	V016	nombre de W.C.
A7	V017	salle de bain
A9	V045	éclairage du logement
A9	V046	humidité du logement
A9	V047	bruit dans le logement
A9	V048	odeurs dans le logement
A10	V049	réparations sanitaires
A10	V050	réparation-amélioration électrique
A10	V051	réparation-amélioration isolation
A10	V052	réparation-amélioration sécurité
A10	V053	ajout de pièces
A10	V054	pose de revêtement de sol,mur
A10	V055	travaux de peinture
A10	V056	réparation extérieure
A11	V063	valeur du logement,minimum
A11	V064	valeur du logement,maximum
A12	V065	raccord gaz
A12	V066	raccord électricité
A12	V067	raccord eau
A12	V068	raccord télédistribution
A13	V069	chauffage central électrique
A13	V070	chauffage central mazout
A13	V071	chauffage central gaz
A13	V072	chauffage central convertible
A13	V073	poële à mazout
A13	V074	poële à charbon-bois
A13	V075	poële au gaz
A13	V076	chauffage-appareil électrique
A14	V077	coût logement mensuel
A15	V078	difficultés ou emprunt loyer
A15	V079	difficultés payer eau,gaz,électricité
A15	V080	difficultés payer chauffage
D3	V084	type d'habitation
D4	V085	évaluation logement par enquêteur
D5	V086	évaluation immeuble par enquêteur
B14	V361	résidence secondaire
B14	V362	logements, immeubles,...
B23	V418	gpe, avantages pour payer logement
B23	V419	gpe, avantages pour payer les charges
x	CRITCO	Critère de cohabitation MIR

3) On souhaitera **obtenir des statistiques** rapides sur les éléments de la base.
Par exemple,

combien y a-t-il de variables relatives au logement dans le fichier individuel ? - réponse : 49
combien de variables nouvelles ? - réponse : 183
combien de variables subjectives ? - réponse : 103
etc.

2.4 - Modalités de consultation de la documentation transversale.

Pour l'analyste qui désire travailler sur un domaine précis, la marche à suivre est la suivante.

1 - consulter par l'intermédiaire des bases de données la liste des variables relatives à ce domaine, disponibles dans le(s) fichier(s).
Cette liste devra comporter le champ "TYPE" qui signale si les variables sont originales ou nouvelles (et CONSULTER).

2 - Si cette liste comporte des variables nouvelles, il obtiendra des renseignements supplémentaires sur leur contenu (leur construction) en consultant les documents de traitement de texte.
Les noms des documents à consulter pour obtenir les renseignements sont dans le champ "CONSULTER" de la base de données.

A partir de la base de données, le consultant dispose des informations nécessaires, soit pour entreprendre la sélection des variables qui l'intéressent (nom des variables, localisation dans les fichiers de travail), soit pour poursuivre sa recherche documentaire (variables nouvelles plus largement explicitées dans des documents).

A l'heure actuelle, il est nécessaire que l'analyste soit appuyé par un agent du service "logistique et documentation" pour effectuer cette démarche.
Mais il est envisageable, une fois la documentation totalement élaborée, de lui laisser une autonomie en établissant des programmes de macro-instructions dans la base de données et en guidant les choix possibles par des menus.

III - L'ASPECT LONGITUDINAL

3.1 - Vers un fichier de gestion longitudinale du panel

La masse d'informations allant croissant dans un panel, il n'est plus envisageable de maintenir sous forme de fichier exploitable TOUTE l'information. Les fichiers SPSS sont des fichiers qui occupent un espace informatique trop important et il devient urgent de tenir compte de cette place de stockage. Pour le stockage de l'information, on envisage de généraliser l'utilisation du système de banques de données SQL, où l'information est concentrée dans un espace minimal.

Ce stockage intermédiaire ne doit en aucune façon être une perte de temps pour l'analyste. Pour qu'il en soit ainsi, la création de fichier SPSS à partir d'un choix de variables SQL doit devenir automatique.

Ce travail est déjà en grande partie réalisé par un interface SQL-SPSS et est amélioré de manière à devenir totalement fonctionnel.

D'autre part, pour que le travail de sélection des variables et des individus (ou groupes, ou ménages) soit performant, il devient nécessaire de bien maîtriser les différents éléments qui servent cette sélection.

A partir de la création des variables nouvelles se rapportant à la vague 1985, on peut distinguer trois grands types de variables.

- certaines sont des variables de travail (par exemple, les sommes de revenus)

- d'autres sont des indicateurs de présence ou absence d'information dans un groupe de variables de travail; par exemple, si la personne est retraitée, alors les variables relatives aux pensions, retraites, etc. comportent de l'information.

- d'autres encore sont plutôt des variables de liaison. En particulier, de nombreuses variables ne font que rapporter de l'information d'un enregistrement à un autre enregistrement. Ce report d'information est une procédure lourde qui est illustrée dans le document V.N.I. (page 49 à 51).

La sélection des variables de travail ne pose pas de problème particulier mais elle doit être facilitée par l'utilisation d'indicateurs.

La prise en compte de l'aspect longitudinal va simplement nécessiter la création d'indicateurs signalant la disponibilité de l'information au fil des vagues.

Ce problème se résume à celui du choix des indicateurs.

La sélection des cas à analyser est beaucoup plus problématique.

D'une part, elle est liée aux variables, puisque doivent être sélectionnés les cas qui présentent un intérêt (de l'information utile) en regard de l'analyse projetée.

D'autre part, elle pose, directement en regard des individus, non seulement le problème de leur appartenance à un groupe, à un ménage mais aussi celui de leur changement d'appartenance au fil des vagues.

Cette sélection des cas cumule donc les problèmes liés à la présence de plusieurs niveaux d'analyse (tel individu du ménage A en 1985 est devenu un individu du ménage B en 1987) et les problèmes liés au report d'information d'un enregistrement à l'autre.

En bref, cette sélection manipule les variables de liaison.

Pour guider les choix de l'analyste, il est donc projeté de constituer un fichier qui rassemble ces informations difficiles à manipuler. C'est ce fichier qui est désigné comme "fichier de gestion longitudinale du panel". Celui-ci doit contenir essentiellement les variables de liaison et des indicateurs.

La constitution de ce fichier de même que, pour l'analyste, le choix et la manipulation des variables de travail seront grandement facilités par une nomenclature systématisée des variables.

3.2 - La nomenclature des variables.

Cette nomenclature des variables repose sur le principe d'homogénéisation exposé sous 1.2.

L'homogénéisation, dans son aspect transversal, consiste à adopter un même nom de variable pour une même information, depuis le questionnaire de départ jusqu'aux fichiers de travail. Dans son aspect longitudinal, elle consiste à appliquer le même nom, au fil des vagues, pour une variable qui est reconduite d'une vague à l'autre, en ne faisant varier que l'indication de la vague.

Diverses contraintes guident les choix de nomenclature.

Dans un panel de périodicité annuelle comme le nôtre, le nom de la variable contient la désignation de l'année. Cette

désignation est requise pour disposer des variables de plusieurs vagues dans un même fichier.

Les changements dans l'information susceptibles de se produire d'une vague à l'autre doivent être signalés dans le nom de la variable.

Ces changements sont de plusieurs ordres.

Si la variable est abandonnée, elle disparaît simplement, ce qui n'a pas d'incidence sur la nomenclature.

Si elle est introduite, le système de nomenclature adopté doit pouvoir l'accueillir : par exemple, si on a une numérotation continue de V001 à V990, on s'exclut la possibilité d'ajouter plus de 9 variables. Il est préférable de prévoir une nomenclature démarrant à V0001.

Si des modifications sont introduites dans le contenu de la variable, il est bon qu'elles soient signalées dans son nom : la question a pu être changée dans son libellé exact; elle a pu être incorporée dans un autre ensemble de questions, des possibilités de réponses ont pu être modifiées (par exemple, une échelle en 5 catégories est devenue une échelle en 7 catégories).

La nomenclature tient compte des contraintes informatiques : le nom d'une variable SPSS ne peut dépasser 8 caractères et doit commencer par une lettre.

Ces contraintes sont plus ou moins impératives. Il est clair que la limite de la longueur du nom accepté par les systèmes informatiques en usage est incontournable alors que les modalités d'introduction des diverses modifications apportées dans les questions tolèrent une marge de liberté.

D'autres critères sont des souhaits destinés à faciliter la tâche. Il serait utile que le niveau de la variable apparaisse dans son nom, qu'on puisse différencier les variables originales des variables créées. En vue de la constitution du fichier de gestion longitudinale du panel, il serait souhaitable que les variables de liaison soient manifestes.

Compte tenu des diverses contraintes et souhaits, un système de nomenclature a été élaboré.

Sa réalisation dans une base de données longitudinale des variables, révéla ses faiblesses et ses forces. Il fut changé en un deuxième système, moins systématique, mais plus pratique.

3.2.1 - Nomenclature longitudinale systématique.

Ce premier système de nomenclature se place dans une perspective longitudinale absolue. Il tente de produire des noms de variables qui transportent toute l'information sur l'évolution de la variable au cours du temps. Il est établi à partir du nom de la variable dans la première vague.

La V001 est la première variable de 1985.
Elle aurait dû s'appeler V85001 et elle commence par la lettre V.

Le tableau qui suit présente un schéma des évolutions possibles des variables, qui servira de base pour l'explicitation de la logique adoptée.

	VAGUE 85	VAGUE 86	VAGUE 87	VAGUE 88	VAGUE 89
	sans changement				
1)	V..001	V86001	V87001	V88001	V89001
	Variable supprimée				
2)	V..002	V86002	V87002	-	-
	Variable ajoutée en 86				
3)	-	P86001	AP87001	AP88001	AP89001
	variable ajoutée en 87				
4)	-	-	Q87001	AQ88001	AQ89001
	Variable ajoutée en 88				
5)	-	-	-	R88001	AR89001
	varaible ajoutée en 89				
6)	-	-	-	-	S89001
	Variable modifiée en 86				
7)	V..003	M86003	MP87003	MP88003	MP89003
	variable modifiée en 87				
8)	V..004	V86004	M87004	MQ88004	MQ89004
	Variable éclatée en 86				
9)	V..005	E86005A	EP87005A	EP88005A	EP89005A
	-	E86005B	EP87005B	EP88005B	EP89005B
	-	E86005C	EP87005C	EP88005C	EP89005C

On constate que le suivi de la variable s'articule autour de sa base constituée d'une numérotation continue.
La V86001 est la variable 001 de l'année 86
C'est ce n° d'ordre qui est reconduit d'une vague à l'autre.

Ce système accepte l'ajout de 999 variables par vague, puisque la numérotation reprend à 0 pour les variables introduites postérieurement à l'origine du panel.
La P86001 n'a rien à voir avec la V86001.

Chaque variable mentionne la vague (dans le cas présent l'année) par une date explicite : V86001 est une variable de la vague 1986, P86001 est une variable de la vague 1986.

Par ailleurs, un code est aussi attribué à l'année.

P signale la vague 1986
Q 1987
R 1988
S 1989
T 1990
U 1991
(V a déjà été utilisé au départ pour 1985)
W 1992
X 1993
Y 1994
Z 1995

Au-delà de ces dix années, des choix devront être effectués : soit utiliser d'autres caractères admis par les systèmes informatiques (par exemple §, &, ...), soit utiliser les lettres du début de l'alphabet non utilisées par ailleurs.

Ainsi, la P86001 signifie la variable 001 de 86 et introduite à l'année P=86.
La Q87001 signifie la variable 001 de 1987 et introduite lors de la vague Q (1987).

Ce système permet de repérer la date d'introduction de la modification.

La AP87001 signifie la variable 001
de la vague ..87...
ajoutée A.....
lors de .P.....

De la sorte, en face du nom de la variable, on connaît immédiatement son historique complet.

La MQ89004 signifie la variable 004
de la vague ..89...
modifiée M.....
en 1987 .Q.....

Les lettres de l'alphabet de A à O servent à mentionner le type de modification qu'a subi la variable.
Toutes les lettres ne sont pas affectées :

A signifie "ajoutée"
D signale un Déplacement de la question.
E signale l'Eclatement de la variable en plusieurs variables unitaires. Une lettre (arbitraire), positionnée après l'ensemble du nom, sert à identifier chaque variable.
F mentionne un changement dans la Formulation de la question.

H est mis pour "Hole" = trou, et repère une variable qui n'est pas demandée lors d'une vague mais réapparaît ultérieurement.
N signale un changement de Niveau.
M signifie "Modifiée"

Ce système devient fonctionnel si l'on s'accorde sur quelques principes :

On introduit des règles de priorité pour trancher des cas litigieux. Par exemple, un changement de niveau l'emporte sur un déplacement (il le contient). Il l'emporte aussi sur une modification de la formulation de la question parce qu'il est plus important.

On délimite de façon stricte la signification des lettres, donc des concepts qui définissent les divers changements. Par exemple, la formulation est un changement dans le libellé de la question qui n'affecte pas les valeurs de réponses (exemple : "exercez - vous une activité à côté de votre retraite ?" est devenu "exercez - vous une activité en plus de votre retraite ?"). La modification est un changement qui affecte les valeurs possibles des réponses : Si une question prévoit les réponses sur une échelle "jamais" - "quelquefois" - "souvent" et qu'on y ajoute la possibilité de réponse "très souvent", c'est une modification.

Nous n'entrons pas dans le détail de ces principes puisque ce système de nomenclature a été abandonné.

Toutefois, ce système, mis en pratique pour l'élaboration de la base de données des variables longitudinales (sur 3 vagues), s'est avéré fonctionnel.

Son avantage majeur est le suivi intégral de l'évolution des variables.

Son inconvénient, par rapport à nos souhaits, est qu'il ne mentionne ni le niveau de la variable, ni la qualité des variables de liaison.

Compte tenu des contraintes absolues (limite de la longueur de la variable) et compte tenu de la qualité des informations de ce panel (durée de vie prévisible, nombre de variables, organisation en niveaux, estimation des modifications possibles), il n'a pas été possible de déterminer un système de nomenclature qui informe intégralement sur l'histoire de la variable tout en informant sur son niveau.

D'autre part, à partir de la connaissance réelle de l'état d'évolution (de transformation) des variables sur 3 ans, grace à l'élaboration de cette base de données longitudinales, il a paru plus important d'introduire le niveau de la variable que son historique intégral.

Une autre nomenclature a été générée.

3.2.2 - Nomenclature longitudinale pratique.

On remarquera que le niveau d'analyse n'est pas contenu dans le nom des variables créées pour la première vague.
L'introduction de cette information obligera donc à renommer toutes ces variables dans les fichiers.
Conséquemment, la contrainte liée à cette donnée tombe et la nouvelle nomenclature peut être envisagée plus librement.

D'autre part, le travail effectué sur la base de données longitudinales (à partir de l'ancienne nomenclature) a permis de se faire une idée de l'importance des modifications des variables.

Les modifications mineures de tous types (formulation, éclatement, ...), c'est à dire celles qui légitiment une nomenclature intégrale du point de vue de l'historique de la variable, ne sont pas très nombreuses.

La modification principale, lors du passage 1985 - 1986, consiste en l'introduction de 647 variables.

Ces ajouts appartiennent eux-mêmes à deux blocs principaux :

- Toutes les variables invoquant une périodicité mensuelle (en particulier les différents revenus), qui étaient demandées pour quatre mois en 1985, le sont pour douze mois en 1986.

- L'année 1985 étant la première année, n'introduit pas de variable servant à accrocher les informations d'une vague à la vague qui la précède. En 1986, de nombreuses variables de ce type apparaissent.
Ces variables sont en fait des variables de liaison.

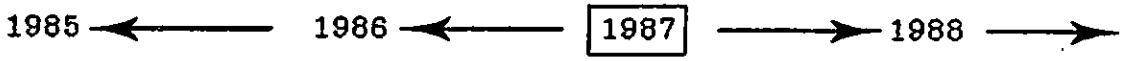
D'autre part, on constate que les changements qui se produisent en-dehors des suppressions simples ou des ajouts simples s'effectuent lors du passage 1985 - 1986 en ce qui concerne les variables de travail. Ils se produisent lors du passage 1986 - 1987 pour les variables de liaison.
En d'autres termes, la deuxième vague est assez stable pour les variables de travail; on peut penser que la troisième l'est assez pour les variables de liaison. Celà s'explique simplement par le fait qu'en 1986 ont été redressées les imperfections de 1985; mais les variables de liaison, qui ne sont introduites qu'en 1986, ne peuvent être améliorées qu'en 1987.

Compte tenu de ces éléments, il a été décidé de faire de 1986 la vague de référence pour la nomenclature des variables de travail et de prendre 1987 comme référence pour les variables de liaison.

Le report des noms d'une vague à l'autre s'effectue donc en amont et en aval de la vague de référence pour ces deux grands types de variables.

Variables de travail

Variables de liaison



Les principes de construction des noms sont simples.

Le premier caractère est une lettre indicatrice du niveau
M..... pour ménage
G..... pour groupe de revenu
I..... pour individu.

Les deux caractères qui suivent mentionnent la vague :
.85..... pour vague 1985
.86..... pour vague 1986, etc.

Si ce sont des variables de liaison, suit l'indication :
...L.... pour liaison
et un numéro d'ordre à 2 positions.

Si ce sont des variables de travail, suit simplement
un numéro d'ordre à 3 positions.

Ces numéros d'ordre commencent à 001 (ou L01) pour chacun des trois niveaux.

Sans compter les variables de liaison, la nomenclature peut ainsi accueillir 999 variables par niveau.

La sécurité est suffisante pour que la désignation des variables introduites ultérieurement se fasse par incrémentation du numéro.

Les changements mineurs qui affectent les variables sont signifiés, au moment où ils se produisent, par une lettre placée après le nom.

Exemples :

M86005 : variable n° 005 de 1986 du niveau Ménage.

I87315 : variable n° 315 de 1987 du niveau Individuel.

G86L04 : variable de liaison n° 04 de 1986 du niveau Groupe.

G86120M : variable 120 de 1986 du niveau groupe, qui a été modifiée en 86 (par rapport à 85).

Si une variable est éclatée pour en constituer plusieurs, les nouvelles variables sont désignées par une lettre qui suit le nom reconduit :

M85110	M86110A	M87110A
-	M86110B	M87110B
-	M86110C	M87110C

Cette lettre sera nécessairement reconduite pour les vagues ultérieures.

En revanche, si la lettre qui suit le nom de base à 6 positions signale une modification d'un autre type, elle n'est pas reconduite :

G85120	G86120F	G87120

On s'accorde sur le fait que la variable de 1987 reste identique à celle de 1986.

Dans ce cas, le fait de laisser tomber cette dernière lettre présente deux avantages :
- on évite l'accumulation des lettres en fin de nom, la taille étant limitée à 8 caractères.
- on repère l'année où la modification est introduite.

Toutefois, il est clair que les indications en fin de nom ne retracent pas l'histoire de la variable. Elles agissent plutôt, pour l'utilisateur, comme un "signal", signifiant qu'il s'est passé quelque chose pour cette variable, mais sans contenir toujours une information sur la nature exacte du changement.

En face d'un nom comme "G86120F" on ne peut pas définir, sans apport d'information externe, s'il s'agit d'une variable dont la formulation a changé entre 1985 et 1986, ou s'il s'agit d'un éclatement de la variable de 1985 en au moins 6 variables (G86120A, ...B, ...C, ...D, ...E, ...F).

Cette nomenclature est donc moins parfaite que la précédente, étant moins systématique.

Toutefois, elle présente un caractère plus pratique, compte tenu du fait que les modifications signalées par des lettres en fin de nom ne sont pas nombreuses.

Elle est aussi plus simple puisqu'elle n'intègre aucun codage artificiel du type "P signifie 86".

C'est donc cette nomenclature qui est en usage dans la base des variables longitudinales qui est présentée dans le paragraphe suivant.

3.3 - La base des données longitudinales.

La base des données longitudinales sert, comme son nom l'indique, à établir les liens entre les différentes vagues.

Cette base est en principe une base des variables. Mais, comme la variable ne doit être que le reflet d'une question du questionnaire, c'est aussi une base des questions.
Elle mentionne également les différents noms intermédiaires qui ont pu rendre compte des questions lors des processus de transformation des observations en données. Ces informations sont nécessaires si l'on veut qu'il n'y ait aucune rupture au fil de ce processus.

C'est lors de la réalisation de cette base que la nécessité s'est fait sentir d'uniformiser totalement les nomenclatures.

L'évolution nécessaire est schématisée comme suit :

		SITUATION EFFECTIVE au 31-12-87	SITUATION IDEALE au 31-12-88
vague 85			
enfant/ adulte	question	B24	I85095
	variable SPSS	V525	I85095
vague 86			
enfant	question	C8	I86095
	nom saisie-SQL	C81	I86095
	nom SPSS	en cours	I86095
adulte	question	D7	I86095
	nom saisie-SQL	D71	I86095
	nom SPSS	en cours	I86095
vague 87			
enfant	question	C3	I87095
	nom saisie-SQL	I87095	I87095
	nom SPSS	- prévisible comme	I87095
adulte	question	D20	I87095
	nom saisie-SQL	I87095	I87095
	nom SPSS	- prévisible comme	I87095
vague 88			
enfant	question	en cours	I88095
adulte	question	en cours	I88095

La base de données qui vient d'être constituée reflète cet état de fait.

Elle contient les champs suivants:

NUMORQST85 numéro d'ordre dans le questionnaire 85
NUMORQST86 numéro d'ordre dans le questionnaire 86.
NUMORQST87 numéro d'ordre dans le questionnaire 87.

Ces trois numéros d'ordre servent à trier les informations dans l'ordre où elles apparaissent dans chaque questionnaire.

QUESTION85 numéro de la question 1985
QUESTION86 numéro de la question 1986
QUESTION87 numéro de la question 1987

A l'avenir, les champs QUESTION ne devraient plus être nécessaires (équivalents aux noms SPSS)

NOM85 nom de la variable 1985
NOM86 nom de la variable 1986
NOM87 nom de la variable 1987

Il s'agit du nom défini par la nomenclature présentée.

NOMSAISI86 nom donné à la saisie 1986

HISTOR85 ancien nom de la variable SPSS de 1985

VARLABEL signification de la variable.

NIVEAU niveau de la variable

LIAISON liaison (oui/non)

OBS8586 observation sur le passage 85-86
OBS8687 observation sur le passage 86-87

Cette base de données est indexée sur les NUMORQST85, NUMORQST86 et NUMORQST87 de manière à permettre d'établir des listes selon les ordres différents des questionnaires, en fonction de l'usage qu'on désire en faire.

Cette base intègre les informations des trois années et contient au total 1535 enregistrements.

A titre d'illustration, nous présentons ci-après un extrait d'une liste tirée de la base de données triée sur 1987, puis 1986, puis 1985.

EXTRAIT DE LA BASE DE DONNEES LONGITUDINALES.

Signification de la variable	n° d'ordre dans questionnaire en 1985	en 1986	en 1987
...			
caméra	40	189	191
équipement de camping	44	193	192
machine à tricoter	24	172	193
micro-ordinateur	33	182	194
caravane	43	192	195
compact-disk	0	194	196
dépenses courses pour ménage	75	195	197
factures-virement bancaire	0	0	198
factures par chèques	0	0	199
factures carte crédit	0	0	200
factures mandat postal/bancaire	0	0	201
factures en liquide	0	0	202
difficultés à joindre les 2 bouts	0	0	203
difficultés ou emprunt loyer	72	162	204
difficultés payer eau,gaz,électricité	73	163	205
difficultés payer chauffage	74	164	206
difficultés alimentation	0	0	207
difficultés payer médecin	0	0	208
difficultés payer vêtements	0	0	209
avec revenu, ménage s'en tire...	0	198	210
montant nécessaire pour joindre 2 bouts	0	199	211
temps (en mois) en cas de coup dur	0	0	212
revenu mensuel considéré très mauvais	0	200	213
revenu mensuel considéré mauvais	0	201	214
revenu mensuel considéré assez mauvais	0	202	215
revenu mensuel consid. pas tropmauvais	0	203	216
revenu mensuel considéré bon	0	204	217
revenu mensuel considéré très bon	0	205	218
évaluation subjective montant impôt	0	0	219
connaissance déduc.intérêts immeuble	0	0	220
connaissance déduc.intér.meuble,voiture	0	0	221
connaissance déduc.primes assurances	0	0	222
connaissance déduc.cotis.CMC	0	0	223
connaissance déduc.cotis.épargne-logmt	0	0	224
connaissance déduc.libéralités	0	0	225
connaissance déduc.frais garde enft...	0	0	226
connaissance déduc.actions,parts soc.	0	0	227
...			

EXTRAIT DE LA BASE DE DONNEES LONGITUDINALES.

- suite -

n° de la question en 1985	en 1986	en 1987	NOM de la VARIABLE 1985	1986	1987	nom saisie 1986	ancien nom 1985
...							
A8	A14	A14	M85081	M86081	M87081	A1424	V040
A8	A14	A14	M85085	M86085	M87085	A1428	V044
A8	A14	A14	M85064	M86064	M87064	A147	V024
A8	A14	A14	M85074	M86074	M87074	A1417	V033
A8	A14	A14	M85084	M86084	M87084	A1427	V043
x	A14	A14	x	M86086	M87086	A1429	x
A16	A15	A15	M85087	M86087	M87087	A15	V081
x	x	A16A	x	x	M87154	x	x
x	x	A16B	x	x	M87155	x	x
x	x	A16C	x	x	M87156	x	x
x	x	A16D	x	x	M87157	x	x
x	x	A16E	x	x	M87158	x	x
x	x	A17	x	x	M87159	x	x
A15	A12	A18	M85054	M86054	M87054	A121	V078
A15	A12	A18	M85055	M86055	M87055	A122	V079
A15	A12	A18	M85056	M86056	M87056	A123	V080
x	x	A18	x	x	M87160	x	x
x	x	A18	x	x	M87161	x	x
x	x	A18	x	x	M87162	x	x
x	A17	A19	x	M86090	M87090	A17	x
x	A18	A20	x	M86091	M87091	A18	x
x	x	A21	x	x	M87163	x	x
x	A19	A22	x	M86092	M87092	A191	x
x	A19	A22	x	M86093	M87093	A192	x
x	A19	A22	x	M86094	M87094	A193	x
x	A19	A22	x	M86095	M87095	A194	x
x	A19	A22	x	M86096	M87096	A195	x
x	A19	A22	x	M86097	M87097	A196	x
x	x	A23	x	x	M87164	x	x
x	x	A24	x	x	M87165	x	x
x	x	A24	x	x	M87166	x	x
x	x	A24	x	x	M87167	x	x
x	x	A24	x	x	M87168	x	x
x	x	A24	x	x	M87169	x	x
x	x	A24	x	x	M87170	x	x
x	x	A24	x	x	M87171	x	x
x	x	A24	x	x	M87172	x	x
...							

IV - BILAN et PERSPECTIVES

L'ensemble de la documentation reste encore à l'état d'ébauche. Pour l'heure, il s'agit de rassembler toutes les informations utiles depuis le début de l'entreprise panel.

Les documents réalisés s'articulent autour de deux thèmes principaux:

la reprise d'une documentation intégrale pour la première vague 1985, cette documentation étant composée des descriptions des variables nouvelles et de bases de données sur la totalité des variables.

l'établissement d'une base longitudinale pour gérer l'information dans son ensemble.

Qu'il s'agisse de l'un ou de l'autre de ces deux aspects, il ne paraît pas inutile de répéter à quel point la réalisation de ces travaux s'impose, idéalement, dès le départ d'une entreprise de panel.

Il est à la fois plus simple et plus rapide d'établir une documentation, même transversale, à mesure que les tâches sont réalisées, quand celles-ci sont encore vivantes dans l'esprit des réalisateurs.

De plus, la mise en place d'une documentation longitudinale, si elle n'est pas suivie régulièrement, devient très vite une tâche comportant de grands risques de confusion.

Non seulement cette documentation inter-vagues s'avère une nécessité pour la fusion des informations et par suite, pour les analyses longitudinales, mais elle est porteuse de fruits annexes.

C'est en élaborant cette documentation qu'on prend conscience de la nécessité d'une gestion concertée de l'information.

Comme l'établissement d'une telle documentation sollicite le recours à des compétences diverses, puisqu'il prend en compte les différentes opérations réalisées dans l'élaboration des données, cette réalisation rend à même de diagnostiquer rapidement les diverses faiblesses rencontrées à telle ou telle étape du processus et doit permettre d'assurer la cohérence du fonctionnement en obligeant la libre circulation de l'information.

C'est à partir de cette réalisation qu'on est en mesure de suivre réellement l'évolution d'un panel.

Comme elle se situe d'emblée dans une optique prospective, l'instauration d'une telle gestion devrait permettre d'éviter les erreurs de parcours.

Celles-ci semblent quasi inévitables si l'on songe à la quantité et à la diversité des informations véhiculées.
Délibérément, le présent document s'est limité à aborder le problème de la documentation sous un seul angle, celui des variables, afin de faciliter l'exposé.
Mais les contraintes de la réalité sont plus nombreuses.

L'aspect informatique a été pour ainsi dire passé sous silence : choix des logiciels, considérations sur l'espace disponible, problèmes de compatibilité entre logiciels, avantages et inconvénients d'un choix de micro ou de "macro" informatique, sont des éléments qui accompagnent par nécessité toutes les décisions à prendre.

La documentation a été présentée "du point de vue de l'analyste" qui en est l'utilisateur final. Il n'a aucunement été abordé sous l'angle propre du service "logistique et documentation" : la logique de la saisie de données, par exemple, n'est pas toujours en accord avec la logique de l'analyse.

Aussi, la réalité qui consiste simplement à amasser la somme des informations utiles à l'élaboration de fichiers longitudinaux de grande taille est une tâche plus complexe que ce document ne le laisse paraître.

De plus, les diverses informations doivent être liées.

A l'heure actuelle, une tâche importante reste à entreprendre : l'étude de la meilleure structuration possible des bases.

Trois bases de données transversales (une par niveau) et une base de données longitudinales ont été réalisées.
Il va de soi que ces bases sont liées entre elles puisque les données transversales ont, dans un panel, un devenir longitudinal.
Les logiciels de bases de données relationnelles permettent d'établir sans problème des liaisons entre bases et de rassembler ainsi dans un même ensemble des informations en provenance de bases diverses.

Il reste à étudier la meilleure structuration possible de l'ensemble des documents en dégageant leurs multiples liens.

D'autre part, le panel luxembourgeois est amené à s'enrichir par des analyses comparatives entre différents pays puisque d'ores et déjà, il travaille en étroite collaboration avec le panel lorrain et est impliqué dans un programme européen de recherche en Belgique, aux Pays-Bas et en Irlande.

Les bases de données devront donc être très vite enrichie par la dimension comparative.

Une autre perspective consiste à faciliter l'usage de la documentation, en particulier celui des bases de données, aux différents utilisateurs.

Cette opération nécessite dans un premier temps une vue détaillée sur les différentes utilisations possibles, ensuite, l'intégration de ces différentes demandes dans des programmes et leur présentation dans une structure facile d'emploi.

Sur ce dernier point, la convivialité est requise.

L'enjeu final est que tout utilisateur, qu'il soit néophyte ou utilisateur chevronné des logiciels de base de données, puisse accéder à l'information qu'il désire par ses propres moyens.

Pour ce faire, les différentes possibilités d'utilisation doivent être encadrées par des menus.

C'est la perspective globale qui guide l'ensemble de la démarche actuelle d'élaboration de la documentation.

Pour que cet objectif soit atteint, des travaux annexes doivent être entrepris comme, par exemple, l'établissement d'un fichier de définitions des termes utilisés dans un sens spécifique dans le cadre de l'entreprise panel (le "ménage", le "groupe de revenus", etc.)

Les travaux en cours doivent aussi être achevés : notamment, l'établissement de la nomenclature longitudinale est à poursuivre en ce qui concerne les variables créées par programme.

Quand, d'une part, les divers points requis au niveau de la documentation seront acquis, avec, d'autre part, une automatisation du passage entre lieu de stockage (I.S.Q.L) et lieu de travail (S.P.S.S.X) et avec les possibilités de transferts entre micro et macro-informatique, la documentation par base de données doit devenir le point de départ d'un système-expert de production de fichiers d'analyses.