

Chapitre. 6 : L'organisation des données pour l'analyse

Au terme du chapitre précédent, tous les instruments (tests et questionnaires) ont été administrés dans les écoles, puis les réponses ont été saisies dans des fichiers informatiques.

A ce stade, il existe un choix d'organisation à faire :

- soit travailler tout de suite avec votre logiciel statistique (comme SPSS, STATA, ou SAS) en important les données directement à partir de chaque fichier de saisie
- soit inclure une étape intermédiaire de fusion des fichiers de saisie à l'intérieur d'une même base de données, avant de commencer à utiliser le programme de traitements statistiques.

C'est cette dernière option que nous allons détailler car elle présente plusieurs avantages :

- elle permet un premier contrôle de cohérence des données avant de commencer les analyses
- elle réduit les risques de confusion lors de l'échange de données avec d'autres utilisateurs et chercheurs.

A. L'organisation des données dans une seule base

Le travail de fusion des différents fichiers de saisie dans une seule base de données demande la mise en place d'une structure relationnelle. Une structure de stockage de type relationnel présente un intérêt chaque fois que les données sont " imbriquées ", ce qui est souvent le cas dans les enquêtes en milieu scolaire, où les élèves sont regroupés dans des classes, qui font elles-mêmes partie d'écoles.

C'est la mise en relation, grâce à des champs clés communs, de tables de données de différents niveaux, qui caractérise une base de données relationnelle.

Les données que nous avons collectées portent sur trois niveaux : l'école, la classe et l'élève. Il est donc nécessaire de posséder trois clés (NUMECOLE, NUMCLASSE, et NUMELEVE) pour organiser l'identification :

NUMECOLE (tables de données issues du questionnaire directeur)
Ce champ clé suffit à identifier chaque école.

NUMECOLE, NUMCLASSE (tables de données issues du questionnaire maître)

L'association de ces deux champs permet d'identifier une classe par le numéro de l'école où elle se trouve, et par le numéro de la classe à l'intérieur de l'école.

NUMECOLE , NUMCLASSE et NUMELEVE (tables de données issues du questionnaire élève et des tests aux élèves)

L'association de ces trois champs permet d'identifier un élève par le numéro de l'école, la classe où il (elle) se trouve, et son numéro d'élève dans la classe.

Au moment de l'analyse statistique, on travaille en général sur une table de niveau élève, où les informations de niveau supérieur (classe et école) sont répétées :

**Table pour l'analyse au niveau élève
(une ligne par élève, une couleur par classe),
incluant des variables de niveau classe**

NUMECOLE	NUMCLASS	NUMELEVE	Score début d'année	Score de fin d'année	Fille (1) ou Garçon (0)	Âge de l'élève	Maitre est une femme	Age du maître	Maitre niveau lycée	Maitre niv. Bac ou +	Effectifs de la classe
1	2	1	12	11	1	8	1	24	0	1	55
1	2	2	15	13	0	9	1	24	0	1	55
1	2	3	8	10	0	9	1	24	0	1	55
1	2	4	3	6	1	8	1	24	0	1	55
1	2	5	16	11	1	8	1	24	0	1	55
2	2	1	4	6	1	8	0	49	0	0	70
2	2	2	11	12	0	7	0	49	0	0	70
2	2	3	10	15	0	9	0	49	0	0	70
2	2	4	5	3	1	8	0	49	0	0	70
2	2	5	9	11	0	8	0	49	0	0	70
3	2	1	18	19	1	8	1	33	1	0	44
3	2	2	10	9	0	7	1	33	1	0	44
3	2	3	13	16	0	8	1	33	1	0	44
3	2	4	12	10	1	9	1	33	1	0	44
3	2	5	6	3	1	8	1	33	1	0	44

Cette structure est à réserver à l'analyse, car elle gaspille beaucoup de place (répétitions dans la zone encadrée). Pour le stockage, mieux vaut garder les variables de niveau différent dans des tables séparées, comme ci-dessous :

NUMECOLE	NUMCLASS	NUMELEVE	Score début d'année	Score de fin d'année	Fille (1) ou Garçon (0)	Age de l'élève
1	2	1	12	11	1	8
1	2	2	15	13	0	9
1	2	3	8	10	0	9
1	2	4	3	6	1	8
1	2	5	16	11	1	8
2	2	1	4	6	1	8
2	2	2	11	12	0	7
2	2	3	10	15	0	9
2	2	4	5	3	1	8
2	2	5	9	11	0	8
3	2	1	18	19	1	8
3	2	2	10	9	0	7
3	2	3	13	16	0	8
3	2	4	12	10	1	9
3	2	5	6	3	1	8

Table de données de niveau élève
(une ligne par élève)

NUMECOLE	NUMCLASS	Maître est une femme	Age du maître	Maître niveau lycée	Maître niv. Bac ou +	Effectifs de la classe
1	2	1	24	0	1	55
2	2	0	49	0	0	70
3	2	1	33	1	0	44

Table de données de niveau classe
(une ligne par classe)

La correspondance entre les données des deux tables se fait à partir des deux champs clés qui leur sont communs (lien relationnel dit « un à plusieurs », puisque les données relatives à un maître, ou à une classe, sont valables pour plusieurs élèves, à savoir tous ceux qui ont été échantillonnés dans cette classe).

Le gain de place, qui semble minime dans l'exemple, prend tout son sens dans la pratique : en effet, lors des premières enquêtes PASEC, le questionnaire maître, d'une douzaine de pages, générerait environ 900 variables. Il vaut donc mieux laisser ces données dans leurs tables de niveau classe d'origine, plutôt que de les dupliquer en 20 exemplaires dans le cas d'un tirage au sort de 20 élèves par classe.

La construction d'une table de données " d'analyse " (qui inclut sur une même ligne des variables de niveau différent) est donc à réserver au tout dernier moment de la préparation des données, quand le nombre de variables construites et sélectionnées à chaque niveau sera devenu plus raisonnable (une vingtaine, et pas 900, dans le cas des variables de niveau maître).

En attendant, le mieux est d'importer les différentes tables de données issues des fichiers de saisie au sein d'une base de donnée relationnelle, en conser-

vant leur intégrité, et en assurant la correspondance entre données élèves, données maîtres, et données niveau école, grâce à des champs clés communs.

Si l'on reprend l'exemple des enquêtes PASEC Côte d'Ivoire de 1995/1996 (seulement pour le CM1), nous avons quatre fichiers ACCESS “ source ” :

- Le fichier de saisie des données des cahiers d'élèves de début d'année (questionnaire aux élèves, test de français, test de mathématiques)
- Le fichier de saisie des données des cahiers d'élèves de fin d'année (test de français et test de mathématiques)
- Le fichier de saisie des données issues du questionnaire aux maîtres
- Le fichier de saisie des données issues du questionnaire aux directeurs

Sauvegardez !

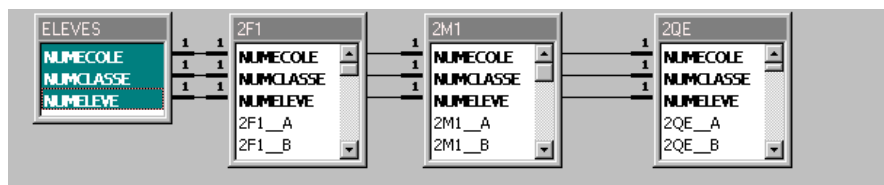
Avant d'aller plus loin, vous devriez vous assurer de la conservation des fichiers originaux de saisie. Concrètement, assurez-vous d'avoir chaque fichier de saisie original en trois exemplaires :

- *Un exemplaire de travail, sur votre disque dur, dans un répertoire au nom explicite (par exemple “ fichiers de travail Côte d'Ivoire 1995-1996 ”*
- *Un exemplaire de sauvegarde, toujours sur votre disque dur, mais à un autre emplacement, par exemple dans un répertoire nommé “ fichiers sources Côte d'Ivoire 1995-96 ”*
- *Un deuxième exemplaire de sauvegarde, cette fois sur un support externe à votre ordinateur. Si la taille de vos fichiers de saisie excède la capacité des disquettes classiques de sauvegarde, sachez qu'il existe de plus en plus de solutions techniques pour faire cette sauvegarde externe (transfert via un réseau ou par câble, lecteur amovible à haute capacité, graveur de CD-ROM, etc.).*

Les manipulations que nous allons proposer maintenant doivent évidemment s'effectuer sur les fichiers de travail. En cas d'altération de ces derniers au cours des manipulations, il vous sera facile de faire une nouvelle copie à partir du double présent sur votre disque dur, et en cas de panne de votre ordinateur, vous pourrez toujours recourir à la sauvegarde externe.

Examinons de plus près le contenu de ces fichiers ACCESS de saisie :

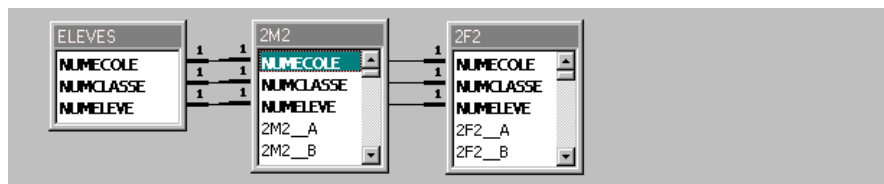
Tables du fichier “ Saisie du cahier de l'élève de début d'année ”



Cette base de données ACCESS comprend quatre tables :

- Une table “ ELEVES ” pour la liste des élèves et de leurs noms
- Une table “ 2QE ” pour les réponses au questionnaire élève
- Une table “ 2F1 ” pour les réponses au test de français de début d'année
- Une table “ 2M1 ” pour les réponses au test de mathématiques de début d'année

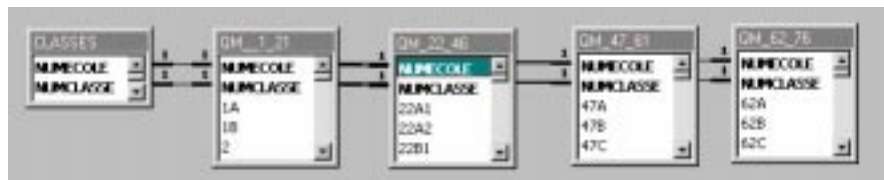
Tables du fichier “ Saisie du cahier de l'élève de fin d'année ”



Cette base de donnée ACCESS comprend trois tables :

- une table “ ELEVES ” pour la liste des élèves et de leurs noms
- une table “ 2F2 ” pour les réponses au test de français de fin d'année
- une table “ 2M2 ” pour les réponses au test de mathématiques de fin d'année

Tables du fichier “ Saisie du questionnaire maîtres ”



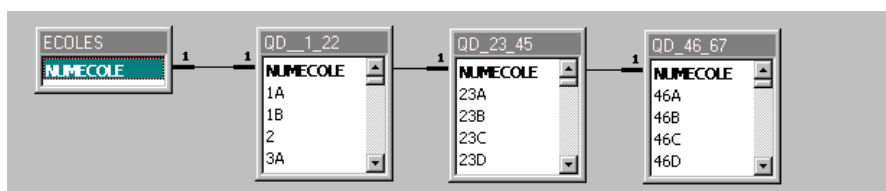
Cette base de données ACCESS comprend cinq tables :

- une table “ CLASSES ” pour la liste des classes et de leurs noms



- une table QM__1_21 pour les questions 1 à 21 du questionnaire maître
- une table QM_22_46 pour les questions 22 à 46 du questionnaire maître
- une table QM_47_61 pour les questions 47 à 61 du questionnaire maître
- une table QM_62_76 pour les questions 62 à 76 du questionnaire maître

Fichier “ Saisie du questionnaire directeurs ”



Cette base de donnée ACCESS comprend trois tables :

- une table “ ECOLES ” pour la liste des écoles et de leurs noms
- une table “ QD__1_21 ” pour les questions 1 à 21 du questionnaire aux directeurs
- une table “ QD_22_45 ” pour les questions 22 à 45 du questionnaire aux directeurs
- une table “ QD_46_67 ” pour les questions 46 à 67 du questionnaire aux directeurs

Chacune de ces bases de données a des tables de même niveau. Les données correspondant au même élève sont donc mises en correspondance dans les différentes tables grâce à des liens du type “ Un à Un ” dont nous avons parlé au chapitre précédent, de même pour les tables de niveau maîtres entre elles, et pour celles de niveau directeur.

L’objectif est maintenant de rassembler toutes ces tables dans une même base de données, c’est-à-dire sous la forme d’un seul fichier informatique (au format ACCESS dans notre exemple).

Il nous faut donc importer chacune des tables citées plus haut dans une nouvelle base de données.

Plusieurs stratégies sont possibles, celle que nous proposons consiste à partir du fichier «Saisie du questionnaire directeurs», d’en faire un double, de le renommer (par exemple «CI_9596.mdb», et d’y ajouter successivement les données du fichier «Saisie du Questionnaire maîtres», puis «Saisie du cahier

de l'élève de début d'année», puis «Saisie du cahier de l'élève de fin d'année».

Reprenons dans l'ordre :

***1. Fusion des données du fichier «Saisie du questionnaire maîtres»
aux données du fichier «CI_9596.mdb»***

Nous supposons que le fichier «Saisie du questionnaire directeurs a déjà été copié et renommé «CI_9596.mdb».

Une fois chargé ce fichier ouvert dans ACCESS, utilisez le menu «Fichier», «données externes», «importer», pour indiquer l'emplacement du fichier «Saisie du questionnaire maîtres» et visualiser son contenu. Importez alors chacune des cinq tables qui font partie de ce fichier (CLASSES, QM__1_21, QM_22_46, QM_47_61, et QM_62_76).

Toutes les tables de niveau Ecole et toutes les tables de niveau Classe sont donc maintenant rassemblées dans la même base. Mais elles ne sont pas encore reliées entre elles par une structure relationnelle.

Pour cela, allons dans la zone de travail «Relations» (par le menu outils), où existent déjà, reliées logiquement entre elles, les quatre tables issues de la saisie du questionnaire directeur.

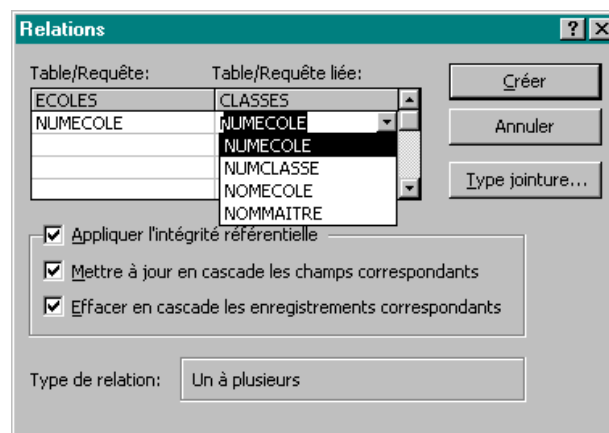
Visualisons les cinq nouvelle table issues du questionnaire maître dans cette zone de travail grâce au menu «Relations», «Ajouter une table». Les liens de type «Un à Un» établis entre ces cinq tables ayant été perdus au cours de l'importation de tout à l'heure, le premier travail consiste à les rétablir comme cela a été décrit au chapitre précédent.

Cela fait, le lien relationnel de type «un à multiple» va être établi entre la table «ECOLES», qui liste les écoles, et la table «CLASSES», qui liste les classes¹.

¹ Rappelons que dans notre exemple, ce lien un à multiple ne s'impose pas, puisque nous n'avons qu'une classe considérée par école, mais nous avons tenu à pouvoir généraliser à une situation où plusieurs classes, ou maîtres, sont échantillonnés par école

Pour cela, il faut reprendre la même technique de création de liens entre tables. La différence vient du fait que cette fois ci, ce sont des tables de niveau différent (Ecole et Classe) qui vont être reliées.

Il suffit pour cela de sélectionner le champ NUMECOLE dans la table ECOLES, et de le faire glisser dans la table CLASSES. Le menu qui s'ouvre alors doit être complété comme l'indique la figure ci-dessous :



Seul le champ NUMECOLE est pris en compte dans la table CLASSES, le champ NUMCLASSE n'étant pas sollicité. Cela suffit pour que le lien entre les deux tables soit du type «Un à Multiple» (il peut y avoir plusieurs classes dans une école).

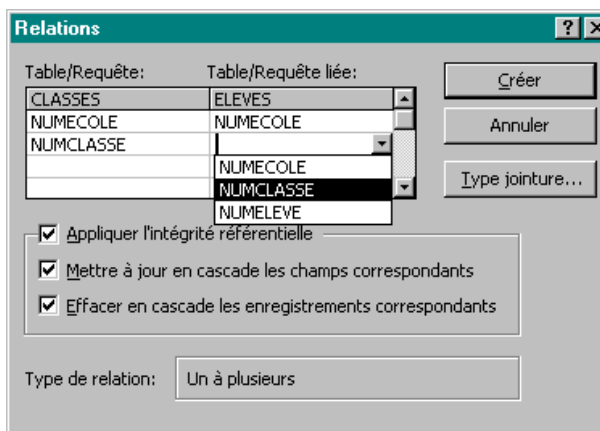
Dans le sous menu «type de jointure», choisissez l'option 2 «Inclure tous les enregistrements de la table ECOLES, et seulement ceux de la table CLASSES pour laquelle les champs joints sont égaux».

Si après avoir cliqué sur «Créer» un message d'erreur s'affiche, c'est probablement parce qu'une valeur du champ NUMECOLE dans la table CLASSES n'a pas d'équivalent dans la table ECOLES. Il peut alors s'agir d'une erreur de codage (même école ayant un numéro de code différent dans le fichier de saisie du questionnaire directeurs et dans le fichier de saisie du questionnaire maîtres), et dans ce cas l'erreur doit être rectifiée dans l'une des deux tables, ainsi que sur les l'exemplaire papier d'origine si la faute vient de là. Il peut aussi s'agir du cas où l'un des directeurs n'a pas rempli son questionnaire (refus, absence,...). Dans ce cas, le numéro d'identification de

l'école en question doit être rajouté manuellement à la table ECOLES. Ces deux types d'erreur ne devraient pas se produire puisque tous les codes possibles ont été introduits «a priori» avant la saisie, mais il peut arriver que pour des raisons diverses de nouveaux codes non prévus au départ aient été ajoutés au cours de la saisie.

2. Fusion des données du fichier «Saisie du cahier de l'élève de début d'année» aux données du fichier «CI_9596.mdb»

La technique est la même : importation des quatre tables du fichier «Saisie du cahier de l'élève de début d'année», rétablissement des liens «Un à Un» entre ces quatre tables, puis mise en correspondance des listes. Cette fois, le lien «Un à Multiple» est à établir entre la table «CLASSES» et la table «ELEVES» comme ci-dessous :



Ce sont les deux champs clés NUMECOLE et NUMCLASSE qui suffisent à établir le lien entre chaque élève, son maître, et son école. Pour le type de jointure, c'est toujours l'option 2 «Inclure tous les enregistrements de la table CLASSES, et seulement ceux de la table ELEVES pour laquelle les champs joints sont égaux» qui est à choisir.

Au moment de créer la relation, à nouveau, un message d'erreur peut apparaître si un élève est rattaché à une classe ou à une école qui n'a pas été définie dans la table «CLASSE» ou la table «ECOLE». La correction d'un code erroné, ou l'ajout d'une classe et/ou d'une école dans la table CLASSES ou la table ECOLES permettent alors de créer la relation.

N'oubliez pas que du fait des options prises lors de la création des liens, tout changement de code et toute suppression d'enregistrement se répercutent automatiquement d'une table à une autre. Par exemple, si vous changez la valeur du champ NUMELEVE dans la table ELEVES pour un enregistrement, ce changement sera répercuté dans la table 2QE, 2F1 et 2M1. De même, la suppression d'une école dans la liste ECOLE supprime non seulement les enregistrements correspondants au niveau des tables de questionnaire maître, mais aussi toutes les informations des élèves pour lesquels le champ NUMECOLE contenait le même numéro. Il convient donc de faire très attention pour ne pas altérer les données à ce stade.

3. Fusion des données du fichier «Saisie du cahier de l'élève de fin d'année» aux données du fichier «CI_9596.mdb»

La procédure est la même que pour la saisie du cahier de début d'année à quelques nuances prêt.

La première est qu'une table ELEVES (liste d'élèves) existe déjà dans la base. Il n'est donc pas nécessaire de l'importer à nouveau, et a priori, il suffit d'importer les deux autres tables (2F2 et 2M2) et de les relier à la table ELEVE existante par un lien de type «Un à Un» pour que l'intégration soit terminée.

En fait, il est probable qu'un message d'erreur apparaîtra au moment de la création de ce lien de type «Un à Un», car il est très fréquent que des erreurs de codage d'élève, ou bien l'inclusion accidentelle d'élèves interrogés en fin d'année scolaire alors qu'ils ne faisaient pas partie de l'échantillon interrogé en début d'année, fassent en sorte qu'un élève de la table 2F2 ou de la table 2M2 ne soit pas déclaré dans la table ELEVES.

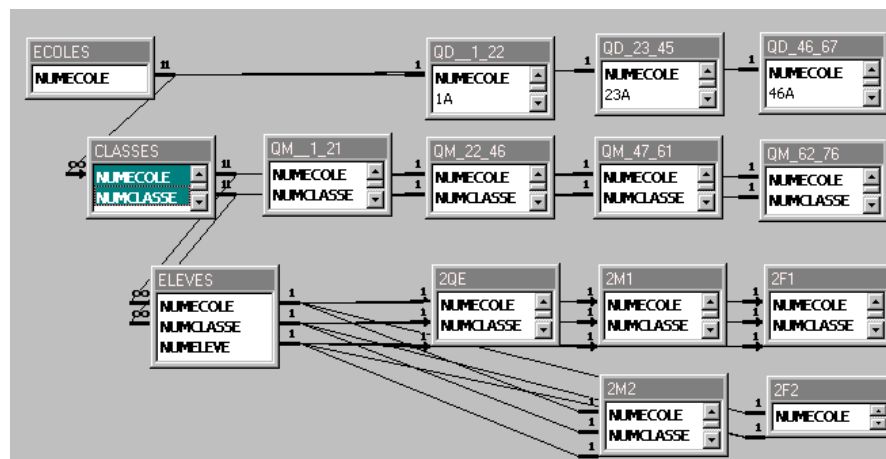
Si c'est le cas, il est alors nécessaire d'importer la table ELEVES issue du fichier «Saisie du cahier de l'élève de fin d'année», et de la comparer avec la table ELEVES issue du fichier «Saisie du cahier de l'élève de début d'année». ACCESS attribue automatiquement un nom différent à cette deuxième table ELEVES lors de l'importation, mais il faudra veiller à la renommer de manière explicite. Une fonction de «recherche de non correspondance entre deux tables», qui existe dans ACCESS, appliquée à la table ELEVES de début d'année en comparaison à la table ELEVES de fin d'année, permettra de repérer ces erreurs de codage ou élèves surnuméraires. Des corrections, qui

nécessitent parfois le retour aux tests et questionnaires originaux, permettront de rectifier les erreurs, et de pouvoir finalement créer le lien. En cas d'ambiguïté du codage (on ne sait pas à quelle école appartient tel ou tel élève), il faut se référer au nom de l'école ou du maître.

Les risques d'altération ou de pertes de données déjà signalés existent également à ce stade, toujours du fait des options de mise à jour et de suppression en cascade définies lors de la création des liens «Un à Un» et «Un à plusieurs». Il est recommandé de tenir un registre des opérations de changement de code ou de destruction effectuées par rapport à la base d'origine, pour éviter d'avoir à tout recommencer à partir de zéro (c'est à dire d'aller rechercher une copie intacte de la base d'origine) en cas de fausse manœuvre.

4. Finalisation de la base de données

A l'issue de toutes ces opérations, le graphique de visualisation des liens de la base «CI_9596.mdb peut être ordonné ainsi :



Ce graphique fait bien apparaître la structure hiérarchisée de notre base, ordonnée autour des trois listes pivot ECOLES, CLASSES, et ELEVES, reliées par des liens «Un à Multiple» (une école comprend plusieurs classes, et une classe comprend plusieurs élèves). Toute caractéristique d'un élève peut donc être reliée aux caractéristiques de sa classe ou de son maître, ainsi que de son directeur ou de son école.

Certes, il aurait été possible de ne pas définir les liens, et de se contenter de rassembler les différentes tables dans un même fichier. Mais l'établissement des liens nous permet de détecter des problèmes éventuels de codage ou de cohérence.

Ce travail de «nettoyage» doit être également complété pour éviter les enregistrements vides dans les tables, source de confusion lors de l'analyse. En effet, la construction des fichiers de saisie nous a amenés à déclarer tous les codes d'école et de classe possibles.

Le cas le plus simple est celui des codes inutilisés (par exemple les numéros de 1 à 120 ont été déclarés, et seulement 100 écoles ont été finalement touchées par l'enquête). Dans ce cas, il suffit de détruire les lignes correspondant à ces codes inutilisés dans la table «ECOLES» (tout simplement en ouvrant la table, en sélectionnant la ligne, en appuyant sur la touche «effacer», et en recommençant autant de fois que nécessaire), et les enregistrements correspondants des tables relatives aux questionnaires directeurs et aux questionnaires maîtres seront détruits en cascade.

Le cas des maîtres ou des directeurs n'ayant pas répondu à leur questionnaire est plus délicat. Il faut détruire les lignes correspondantes dans les tables qui contiennent les réponses aux questionnaires (par exemple, dans 2QM__1_21, 2QM_22_46, 2QM_47_61 et 2QM_62_76, s'il s'agit d'un maître qui n'a pas répondu), mais ne pas détruire les enregistrements correspondant à ces maîtres ou à ces écoles dans les tables «ECOLES» et «CLASSES», sous peine de voir disparaître les données relatives aux tests et aux questionnaires des élèves de ces maîtres ou de ces directeurs.

Enfin, il convient de supprimer les tests ou les questionnaires élèves non remplis, dans chacune des tables correspondantes, qui ont dû être signalés par des champs du type «REMPLI» ou «NON_REMPLI». Là encore, il ne faut pas détruire directement l'enregistrement au niveau de la table «ELEVES», sous peine, par exemple, de faire disparaître le test de français d'un élève qui n'aurait été absent que pour le seul test de mathématiques, ou les données du cahier de l'élève de début d'année pour un élève qui aurait été absent lors de l'administration du cahier de fin d'année.

En théorie, la liste «ELEVES» ne doit pas contenir d'enregistrements correspondant à des élèves n'ayant participé ni au test de début d'année ni au test de

fin d'année (si tel était le cas, il faudrait les supprimer au niveau de cette même table «ELEVES»). Par contre, il peut arriver que des élèves qui n'avaient pas participé à l'administration de début d'année aient eu à remplir un cahier de l'élève de fin d'année. S'il s'agit d'élèves isolés dans une classe, c'est une erreur d'administration des tests en fin d'année, et il convient de supprimer les enregistrements correspondants dans la table «ELEVES». Par contre, il peut s'agir du cas ou une école de l'échantillon qui n'était pas accessible en début d'année est devenue accessible en fin d'année. Dans ce cas, il convient de conserver les informations relatives à ces élèves, ce qui permettrait, notamment, en restant plus près de l'échantillon théorique, de mieux reconstituer un score moyen pour la population considérée.

Tous ces contrôles ne concernent que deux aspects du nettoyage des données : la vérification de la cohérence du codage (par exemple, le code du champ ECOLE doit être le même pour un élève, pour son maître et pour son directeur), et la suppression des enregistrements vides (correspondant à des questionnaires ou à des tests que l'individu n'a jamais remplis). Le problème des erreurs de saisie, ou des incohérences dans les réponses (par exemple un élève qui aurait à la fois coché «GARCON» et «FILLE») est plus facile à traiter dans le cadre du logiciel statistique que dans celui du logiciel de base de données. Il est donc temps de passer à la phase suivante, celle de l'exportation des données au format d'un logiciel statistique, et de l'élaboration d'une stratégie de manipulation des données dans ce logiciel.

B. Importation et manipulation des données dans un logiciel statistique

Nos données sont maintenant rassemblées au sein d'un fichier unique qui correspond à une base de données relationnelles (au format ACCESS dans notre exemple).

Certains traitements (calculs d'effectifs, pourcentages d'élèves ayant telle ou telle caractéristique) pourraient être effectués en restant dans le cadre d'un logiciel de base de données.

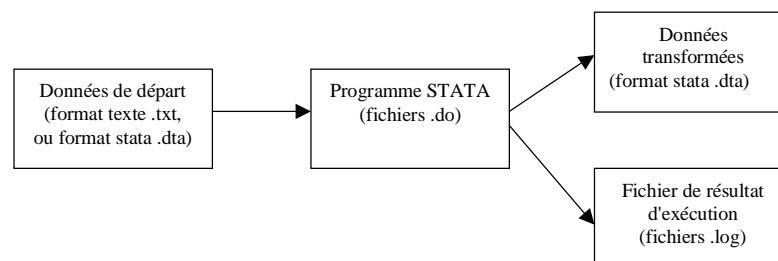
Néanmoins, pour plus d'efficacité, nous suggérons de bien séparer les rôles : saisie et stockage des données pour le logiciel de base des données, et préparation des données et analyses dans le cadre du logiciel statistique.

Le logiciel statistique que nous utilisons dans nos exemples est STATA, mais les mêmes principes sont valables pour SPSS, SAS, etc.

Il existe deux modes de travail avec un logiciel statistique : le mode «interactif» et le mode «programme». Dans le mode interactif, assez semblable à celui de l'utilisation normale d'un tableur, vous effectuez des manipulations diverses (tri, calculs, etc...) directement à l'écran, ce qui a l'avantage de la souplesse. Dans le mode «programme», vous indiquez dans un fichier programme la succession des opérations que vous désirez effectuer, et celles-ci sont ensuite exécutées les unes après les autres, automatiquement.

Nous allons proposer une organisation du travail à partir du mode «programme» pour parvenir jusqu'au stade de l'analyse des données, et ce pour profiter de l'avantage essentiel de ce mode : garder en mémoire toutes les transformations et manipulations effectuées à partir des données brutes, ce qui permet de pouvoir retracer toutes les opérations qui ont permis d'obtenir tel ou tel résultat, tout en facilitant la recherche d'erreurs éventuelles.

Pour cela, il faut comprendre comment fonctionne le mode «programme» d'un logiciel comme STATA :



Chaque rectangle ci-dessus a pour équivalent un fichier informatique. L'exécution du programme STATA, au centre, provoque typiquement un certain nombre d'événements :

- le chargement dans la mémoire de l'ordinateur des données contenues dans un fichier de données de départ (au format texte, avec l'extension .txt, ou au format STATA, avec l'extension .dta)
- l'exécution d'un certain nombre d'opérations (codage de variables, calcul de moyennes, etc.)
- la sauvegarde du résultat de toutes ces opérations dans un fichier de résultat (avec l'extension .log)
- la sauvegarde des données transformées dans un fichier de données (généralement au format stata, avec l'extension .dta)

C'est une généralisation de ce fonctionnement avec plusieurs fichiers programmes que nous vous proposons dans le schéma d'analyse suivant, toujours avec l'exemple des données PASEC de 1995-1996 en Côte d'Ivoire au niveau CP.

Cet exemple peut facilement être reconstitué en recopiant dans votre disque dur le contenu du répertoire du CD-ROM PASEC qui concerne la Côte d'Ivoire. Si vous ne disposez pas de ce CD-ROM, voici comment se présente l'arborescence des répertoires concernant la Côte d'Ivoire, une fois les fichiers recopiés sur le disque dur :

