Projet Obe Maghreb

Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

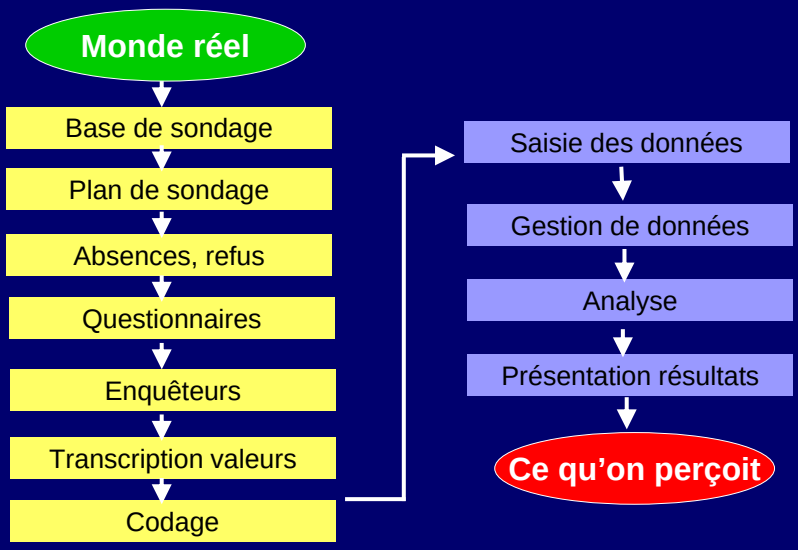
Accès aux données (saisie)

Agnès Gartner, Pierre Traissac
UMR 204 « Prévention des malnutritions et pathologies associées »
IRD, Montpellier, France

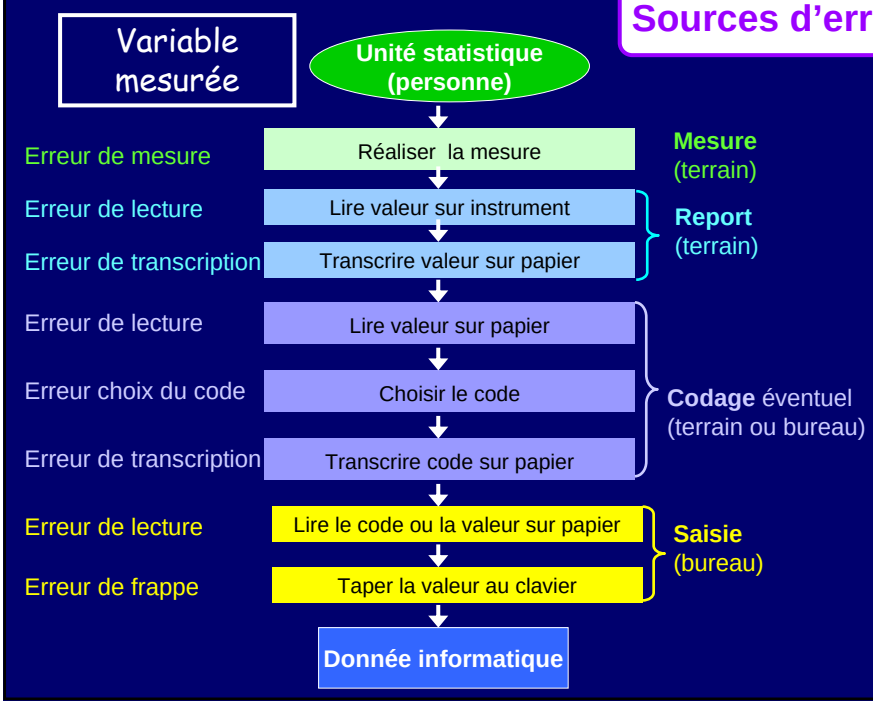
Pourquoi EpiData

1- Saisie des données (Masques de saisie, contrôles, validation)	**1- EpiData**
2- Gestion des données (Recodages, calculs de nouvelles variables, fusion de fichiers, sauvegardes)	**2- EpiData Analysis (+ Anthro)**
3- Analyse des données (statistiques descriptive, inférentielle, tableaux, graphiques)	**3- EpiData Analysis**
4- Présentation des données (tableaux, graphiques, document écrit, présentation orale)	**4- MS Office** (Word, Excel, Power Point) (ou OPEN Office)

Des données à l'information

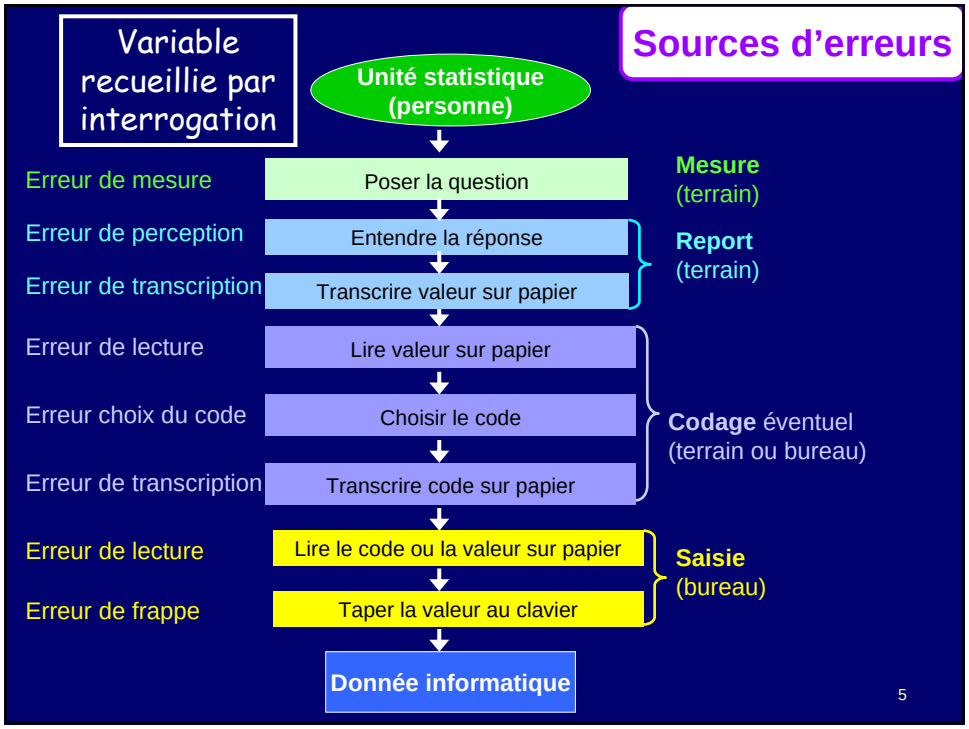


Sources d'erreurs



Sources d'erreurs

Variable recueillie par interrogation



Saisie des données

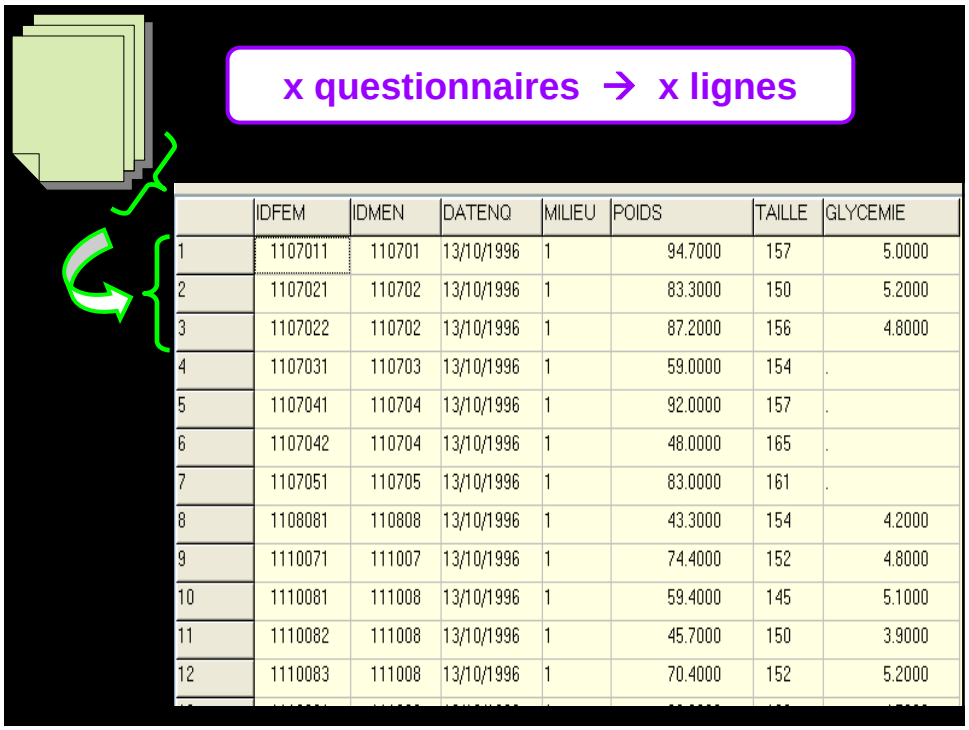
- **Acquisition des données**
  - **saisie = lecture du questionnaire & frappe au clavier**
- **Disposer données sur support informatique**
  - **fidèles à celles recueillies sur terrain**
  - **cohérentes**
  - **documentées** (dictionnaire de variables) **!!!**
- **Recherche et correction d'erreurs**

Détection et correction d'erreurs

- **Erreurs liées à la mesure**
  - réalisation mesure, lecture et transcription
    - formation enquêteurs
  - supervision
    - vérification des questionnaires « le plus près possible » du terrain
- **Erreurs liées au codage**
  - lecture, choix du code, transcription
    - formation codeurs, guide de codage
- **Erreurs de saisie**
  - Erreur de lecture, erreur d'entrée (frappe clavier, choix dans un menu)
    - Ergonomie du questionnaire papier, du masque de saisie
  - Formation opérateurs de saisie

Détection et correction d'erreurs

x questionnaires → x lignes



	IDFEM	IDMEN	DATENQ	MILIEU	POIDS	TAILLE	GLYCEMIE
1	1107011	110701	13/10/1996	1	94.7000	157	5.0000
2	1107021	110702	13/10/1996	1	83.3000	150	5.2000
3	1107022	110702	13/10/1996	1	87.2000	156	4.8000
4	1107031	110703	13/10/1996	1	59.0000	154	.
5	1107041	110704	13/10/1996	1	92.0000	157	.
6	1107042	110704	13/10/1996	1	48.0000	165	.
7	1107051	110705	13/10/1996	1	83.0000	161	.
8	1108081	110808	13/10/1996	1	43.3000	154	4.2000
9	1110071	111007	13/10/1996	1	74.4000	152	4.8000
10	1110081	111008	13/10/1996	1	59.4000	145	5.1000
11	1110082	111008	13/10/1996	1	45.7000	150	3.9000
12	1110083	111008	13/10/1996	1	70.4000	152	5.2000

Saisie des données

- Intégrité d'entité (! identifiants principaux)
- Intégrité de référence (! identifiants secondaires)
- Intégrité de domaine (valeurs de chaque variable)
- Cohérence (2 variables ou plus)
- Autres erreurs (erreurs du 5ème type ☺ …)

identifiants

- Intégrité d'entité (! identifiants principaux)
- Intégrité de référence (! identifiants secondaires)

saisis ou programmés

ménages

	IDMEN	DATENQ	REGION	HABITAT	ECO1	DEPENSES
1	110701	13/10/1996	1	1	74.6000	150
2	110702	13/10/1996	1	1	65.5400	150
3	110703	13/10/1996	1	1	67.2600	200

femmes

	IDFEM	IDMEN	DATENQ	MILIEU	POIDS	TAILLE	GLYCEMIE
1	1107011	110701	13/10/1996	1	94.7000	157	5.0000
2	1107021	110702	13/10/1996	1	83.3000	150	5.2000
3	1107022	110702	13/10/1996	1	87.2000	156	4.8000

domaine

➢ **Intégrité de domaine (valeurs de chaque variable)**

	IDFEM	IDMEN	DATENQ	MILIEU	POIDS	TAILLE	GLYCEMIE
1	1107011	110701	13/10/1996	1	94.7000	157	5.0000
2	1107021	110702	13/10/1996	1	83.3000	150	5.2000
3	1107022	110702	13/10/1996	1	87.2000	156	4.8000

1107011 à 8324061

et UNIQUE

13/01/1996 à 28/12/1997

135 à 185

taille=129 ?

11

Cohérence des questions

(5)	Sexe du CM	Homme ..... 1 Femme 2	—
(6)	Age du CM	Noter l'âge du chef de ménage en année (Si possible vérifiez sur la carte d'identité)	— —
(7)	Ethnie du chef de ménage	Wolof ..... 01 Diola ..... 06 Lébou ..... 02 Mandingue ..... 07 Peuhl ..... 03 Socé/Sarakholé/Soninké .. 08 Toucouleur ..... 04 Mandjaque ..... 09 Sérer ..... 05 Autre ethnie 10	— —
(8)	Sait-il/elle lire ou écrire une phrase dans une langue quelconque ?	Oui ..... 1 Non 2	— —
(9)	Quel niveau scolaire a t-il/elle atteint ?	Pas d'école ..... 0 Brevet diplôme ..... 4 Ecole Coranique ..... 1 Bac diplôme ..... 5 ...upérieur 6	—
(10)	Quelle a été la principa... au cours des 12 dernier...	...er/aide fami. 08 ...giaire ..... 09 ...ant ..... 10 ..... 11 ...tier ..... 12 ...if ..... 13 Pêcheur 07	— —
(11)	État matrimonial du chef de ménage	Marié(e) ..... 1 Veuf(ve) ..... 3 Célibataire ..... 2 Divorcé(e) 4	—
(12)	Si le chef de ménage est un homme	Nombre d'épouses ..... Si femme chef de ménage	—

Recherche sur papier ou en informatique

Détection et correction d'erreurs

- **Détection & correction d'erreurs**
  pendant la saisie :
  contrôles d'entrée

- **Détection & correction d'erreurs**
  après la saisie :
  validation (double saisie)
  apurement

Saisie des données

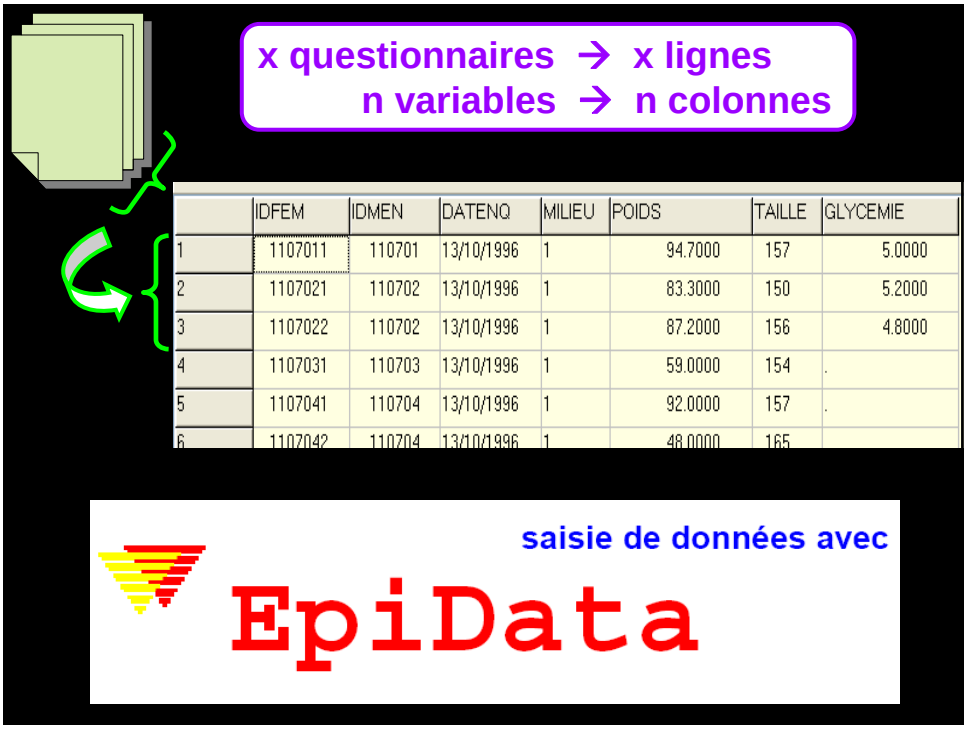
- **Documenter la saisie**
  - **incidents, problèmes pendant la saisie**
  - **dictionnaire de variables**
  - **version des fichiers**
  - **corrections** (par programmation et non mode interactif)
- **!!! Sauvegarde des fichiers !!!**

Logiciel de saisie « data entry »

- **Fonctionnalités spécifiques**
  - **- création « masques de saisie »**
  - **- contrôles d'entrée**
  - **- validation (double saisie)**
  - **- documentation (programmes, codages, …)**

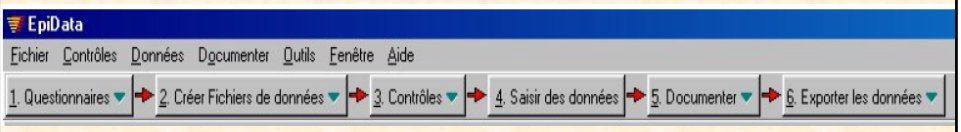
- **Saisie directe interfaces genre tableur (Excel, autres) ou « feuilles de données » logiciel statistique :**
  NON, NON et NON !!!
  (sauf très petit ensemble de données)

15



	IDFEM	IDMEN	DATENQ	MILIEU	POIDS	TAILLE	GLYCEMIE
1	1107011	110701	13/10/1996	1	94.7000	157	5.0000
2	1107021	110702	13/10/1996	1	83.3000	150	5.2000
3	1107022	110702	13/10/1996	1	87.2000	156	4.8000
4	1107031	110703	13/10/1996	1	59.0000	154	.
5	1107041	110704	13/10/1996	1	92.0000	157	.
6	1107042	110704	13/10/1996	1	48.0000	165	

EpiData

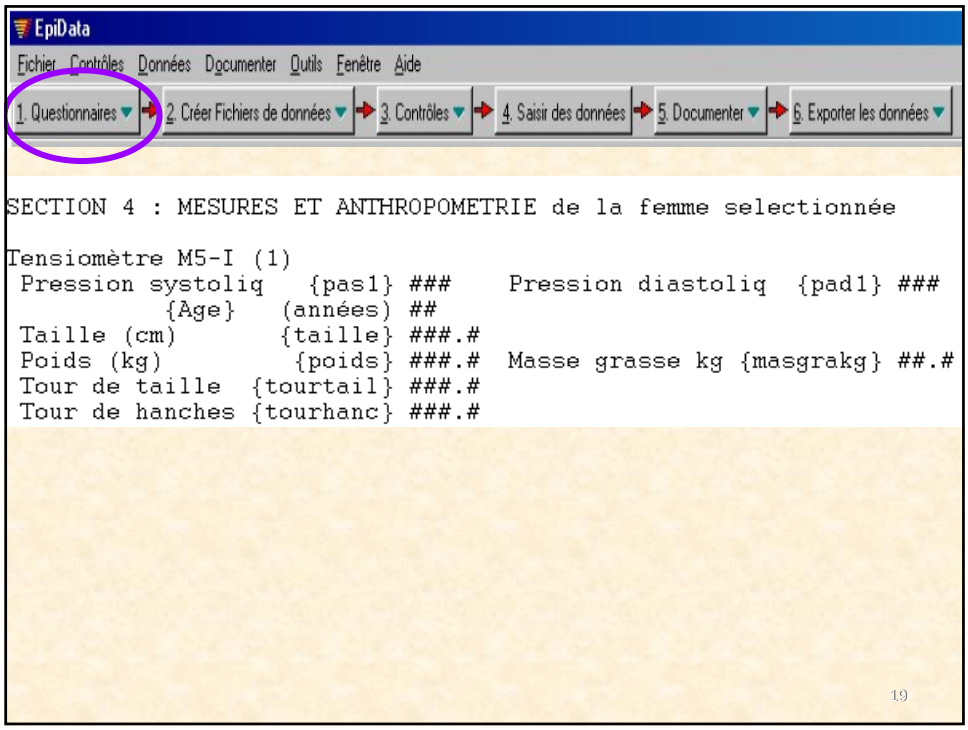


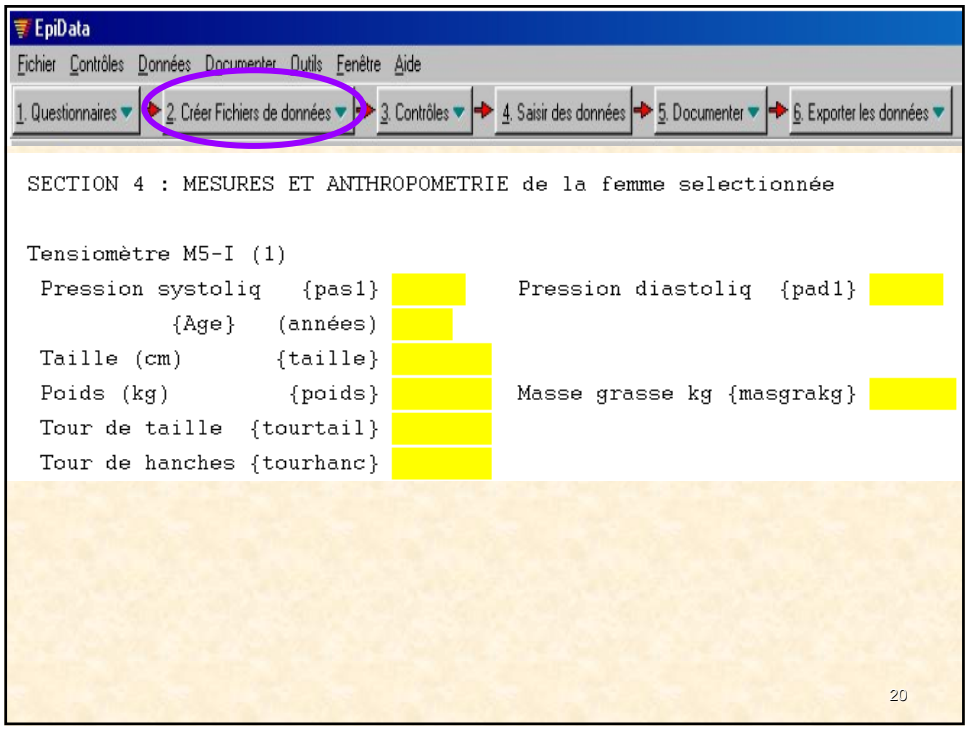
17

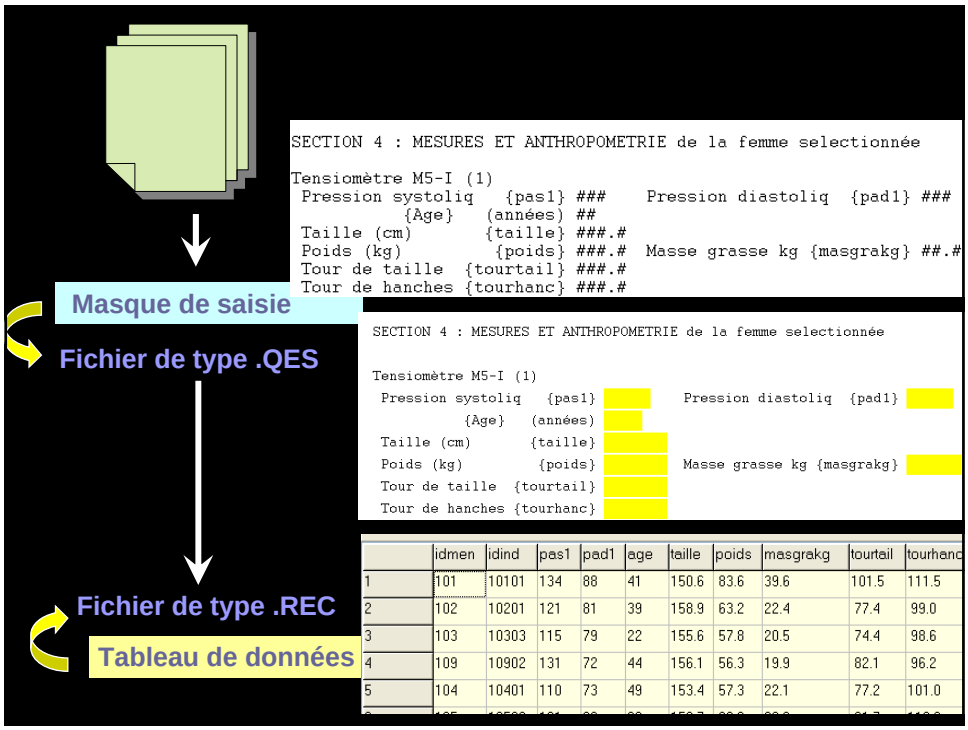
SECTION 4 : **MESURES et ANTHROPOMETRIE** de la femme sélectionnée

Tensiomètre M5-I (1ère mesure)	
Pression systolique |__|__|__|	**Pression diastolique** |__|__|__|
Age (années) |__|__|	
Taille (cm) |__|__|__|,|__|	
Poids (kg) |__|__|__|,|__|	**Masse grasse** (kg) |__|__|,|__|
Tour de taille (cm) |__|__|__|,|__|	
Tour de hanches (cm) |__|__|__|,|__|	

18


EpiData
Fichier Contrôles Données Documenter Outils Fenêtre Aide
1. Questionnaires
2. Créer Fichiers de données
3. Contrôles
4. Saisir des données
5. Documenter
6. Exporter les données
SECTION 4 : MESURES ET ANTHROPOMETRIE de la femme selectionnée
Tensiomètre M5-I (1)
Pression systoliq {pas1} ### Pression diastoliq {pad1} ###
{Age} (années) ##
Taille (cm) {taille} ###.#
Poids (kg) {poids} ###.# Masse grasse kg {masgrakg} ##.#
Tour de taille {tourtail} ###.#
Tour de hanches {tourhanc} ###.#
19


EpiData
Fichier Contrôles Données Documenter Outils Fenêtre Aide
1. Questionnaires
2. Créer Fichiers de données
3. Contrôles
4. Saisir des données
5. Documenter
6. Exporter les données
SECTION 4 : MESURES ET ANTHROPOMETRIE de la femme selectionnée
Tensiomètre M5-I (1)
Pression systoliq {pas1} Pression diastoliq {pad1}
{Age} (années)
Taille (cm) {taille}
Poids (kg) {poids} Masse grasse kg {masgrakg}
Tour de taille {tourtail}
Tour de hanches {tourhanc}
20



Principes et règles de base

- **"Dessiner" l'outil**
  à travers lequel les données seront saisies
- **Créer la liste des variables**
  (colonnes du tableau de données)
- **Pour chaque variable**
  - Choisir un nom
  - Définir le type
  - Définir le format
- **Soigner la présentation**
  - Présentation fidèle à celle du questionnaire papier
    ➡ facilite le travail de saisie
  - Commentaires pour clarifier le questionnaire

Noms des variables

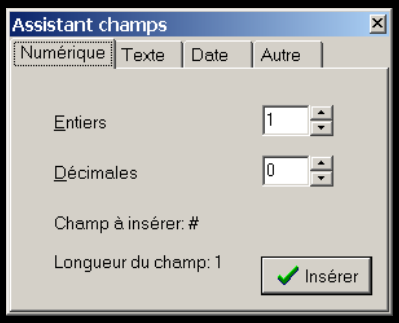
- **Règles à retenir**
  - **- Pas plus de 8 caractères**
  - **- Chiffres autorisés mais pas en 1er caractère**
  - **- Pas de caractères spéciaux (* ; _ , % . & : etc.)**
  - **- Éviter caractères accentués**

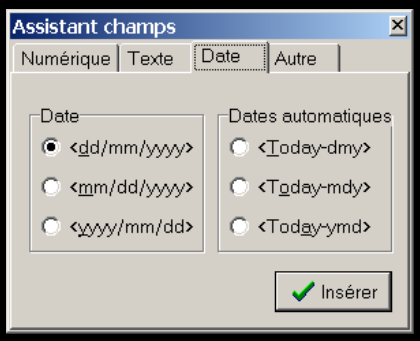
- **Plusieurs façons de les créer**
  - **- Le premier mot de la question**
  - **- Les 8 premiers caractères**
  - **- Les caractères entre { }** (y compris si plusieurs sur la même ligne) **:**
    conseillé

Types et format des variables

- **Alphanumérique (caractère)**
  - **- Standard :** _____ (autant de traits que de caractères possibles)
  - **- Majuscules : <A >** (toutes entrées transformées en majuscules)
- **Numérique**
  - **- Entiers simples :** ### (autant de # que de chiffres voulus)
  - **- Avec décimales :** ##.###
- **Date**
  - **- Européennes : <dd/mm/yyyy>**
  - **- Anglosaxonnes : <mm/dd/yyyy>**
- **Autres**
  - **- Dichotomiques : <Y>** (n'accepte que 'Y' ou 'N', ou '1' ou '0')
  - **- Numéros séquentiels : <idnum>**
  - **- Date du jour : <today-dmy>**
  - **- Cryptées : <S >**

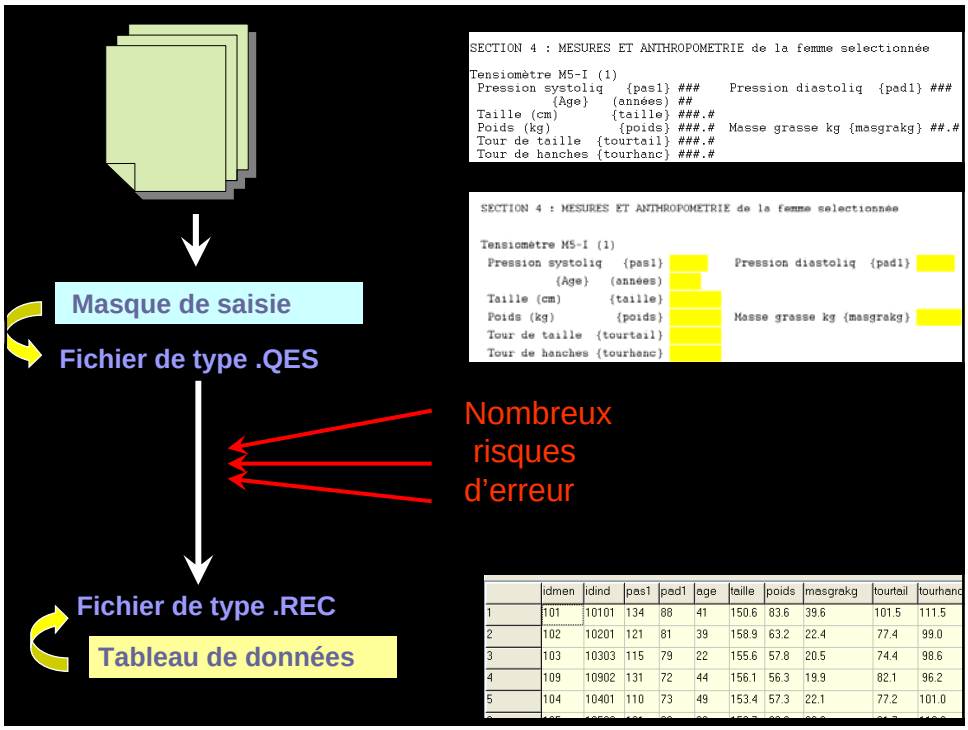
Types et format des variables





Pièges à éviter

- **Mélanger plusieurs u.s. dans un même fichier**
  exemple :
  - le ménage et la personne
  - la personne et ses 'n' hospitalisations
- **Identifiants non uniques, voire manquants !**
- **Oublier les identifiant(s) de niveau supérieur !!**
  nécessaires à la mise en relation hiérarchique des fichiers
  (très classique : « *quel est le ménage de cette personne ??* »)



Détection et correction d'erreurs

- **Détection & correction d'erreurs**
  pendant la saisie :
  contrôles d'entrée

- **Détection & correction d'erreurs**
  après la saisie :
  validation (double saisie)
  apurement

Contrôle de la saisie

- **But :** limiter au maximum les erreurs de saisie,
  tout en facilitant et simplifiant le travail de saisie

- **Création d'un fichier de contrôle** (extension .CHK)
  dans lequel sont spécifiées toutes les opérations à faire :

- **De la plus simple…**
  - Définir les valeurs autorisées pour chaque variable
  - Rendre la saisie obligatoire dans un champ donné
  - Remplir automatiquement certains champs
  - S'assurer de l'unicité des valeurs des identifiants
  - Déplacement conditionnel du curseur (sauts)

- **A la plus compliquée !**
  - Attribuer et faire apparaître des « labels »
  - Utilisation de boîtes de dialogue
  - Faire des calculs cumulatifs
  - Vérifications sur plusieurs fichiers liés etc.

Contrôle de la saisie

- **Principes généraux**

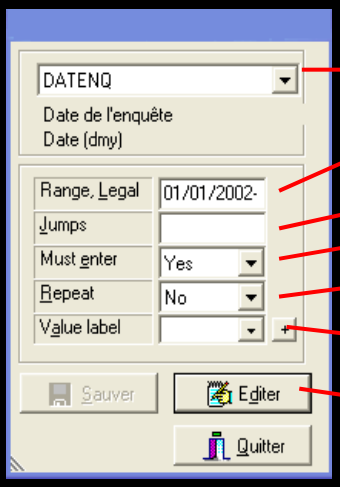
  - Le fichier de contrôle (.CHK) doit porter le même nom que le fichier des données (.REC)

  - L'écriture du fichier de contrôle peut se faire :
    - de façon assistée (boîte de dialogue, pour contrôles simples)
    - en éditant le fichier (contrôles plus complexes, syntaxe)

  - Principes (conseils…)
    - Adapter la complexité des contrôles au type de questionnaire, au niveau des opérateurs de saisie… et à son propre niveau !
    - Commencer par les contrôles simples en utilisant l'assistant, puis éditer le fichier pour les opérations plus complexes
    - Tester les opérations complexes au fur et à mesure

Contrôle de la saisie

- **L'assistant**



Nom de variable + description

Valeurs autorisées: continues (Range) ou non (Legal)

Sauts conditionnels

Champ à entrée obligatoire

Champ à valeur répétée (d'un enregistrement à l'autre)

Attribution de labels: prédéfinis ou création

Pour modifier ou compléter les contrôles (par ex KEY UNIQUE)

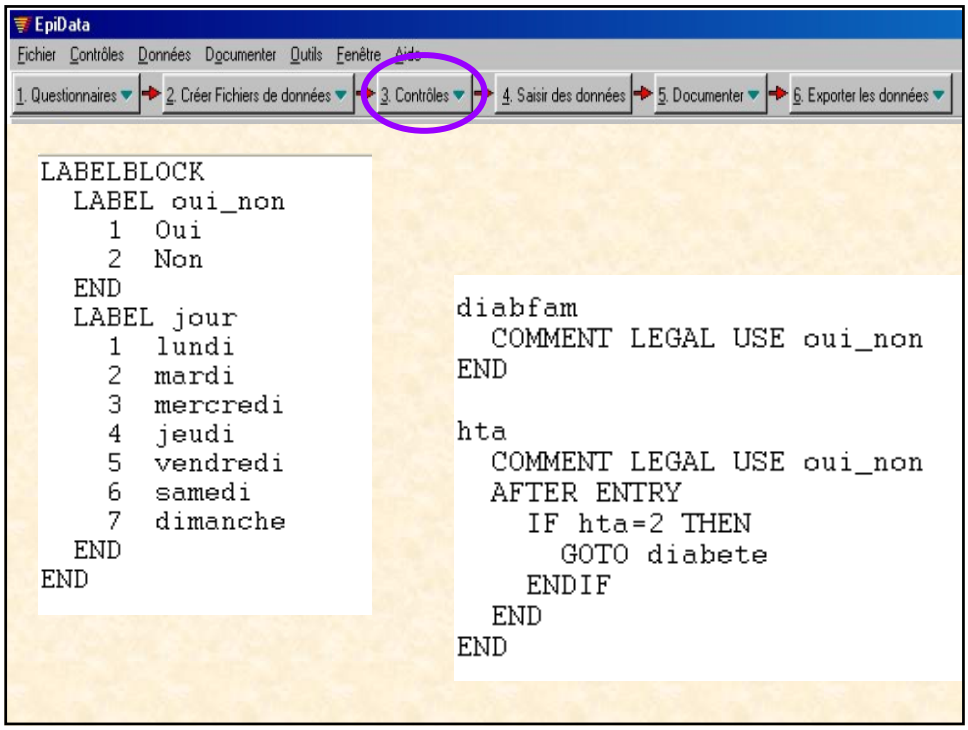
Contrôle de la saisie

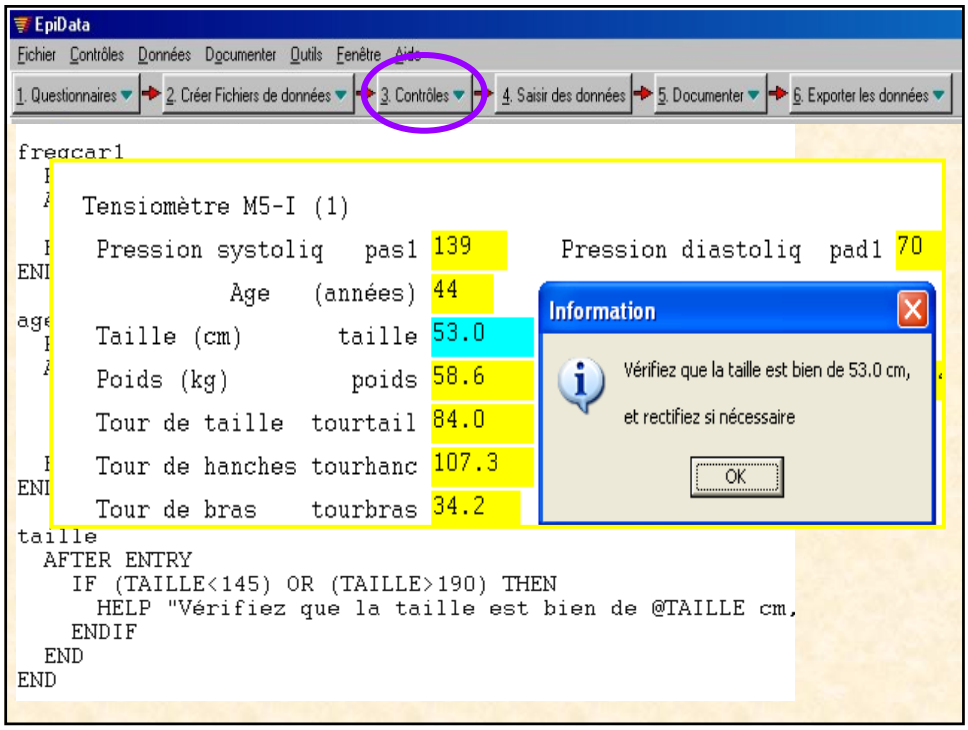
- **Structure**

- La structure du fichier de contrôle comporte une succession de « blocs de commandes », tous terminés par le mot clef END

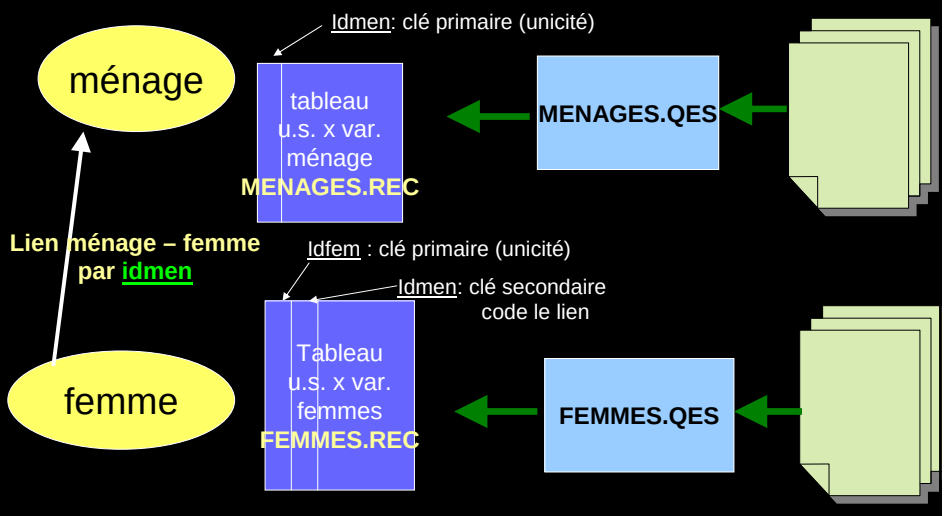
- Les blocs de commande peuvent commencer par:
  - Un NOM DE VARIABLE: les opérations concernent la variable, ou en tout cas seront effectuées lors de la saisie lorsque le curseur se trouvera dans le champ correspondant
  - BEFORE (ou AFTER) FILE (ou RECORD, ou ENTRY): les opérations seront effectuées au début (ou à la fin) de la saisie d'un fichier (ou d'un enregistrement, ou d'un champ)
  - LABEL (ou LABELBLOCK): pour préciser les listes de codes

- Il n'est pas possible d'écrire des instructions entre deux blocs de commande; mais un bloc peut être inclus dans un autre bloc


EpiData
Fichier Contrôles Données Documenter Outils Fenêtre Aide
1. Questionnaires
2. Créer Fichiers de données
3. Contrôles
4. Saisir des données
5. Documenter
6. Exporter les données
LABELBLOCK
LABEL oui_non
1 Oui
2 Non
END
LABEL jour
1 lundi
2 mardi
3 mercredi
4 jeudi
5 vendredi
6 samedi
7 dimanche
END
END
diabfam
COMMENT LEGAL USE oui_non
END
hta
COMMENT LEGAL USE oui_non
AFTER ENTRY
IF hta=2 THEN
GOTO diabete
ENDIF
END
END


EpiData
Fichier Contrôles Données Documenter Outils Fenêtre Aide
1. Questionnaires
2. Créer Fichiers de données
3. Contrôles
4. Saisir des données
5. Documenter
6. Exporter les données
fregcar1
Tensiomètre M5-I (1)
Pression systoliq pas1 139
Pression diastoliq pad1 70
Age (années) 44
Taille (cm) taille 53.0
Poids (kg) poids 58.6
Tour de taille tourtail 84.0
Tour de hanches tourhanc 107.3
Tour de bras tourbras 34.2
Information
Vérifiez que la taille est bien de 53.0 cm,
et rectifiez si nécessaire
OK
taille
AFTER ENTRY
IF (TAILLE<145) OR (TAILLE>190) THEN
HELP "Vérifiez que la taille est bien de @TAILLE cm,
ENDIF
END
END

Modèle de données

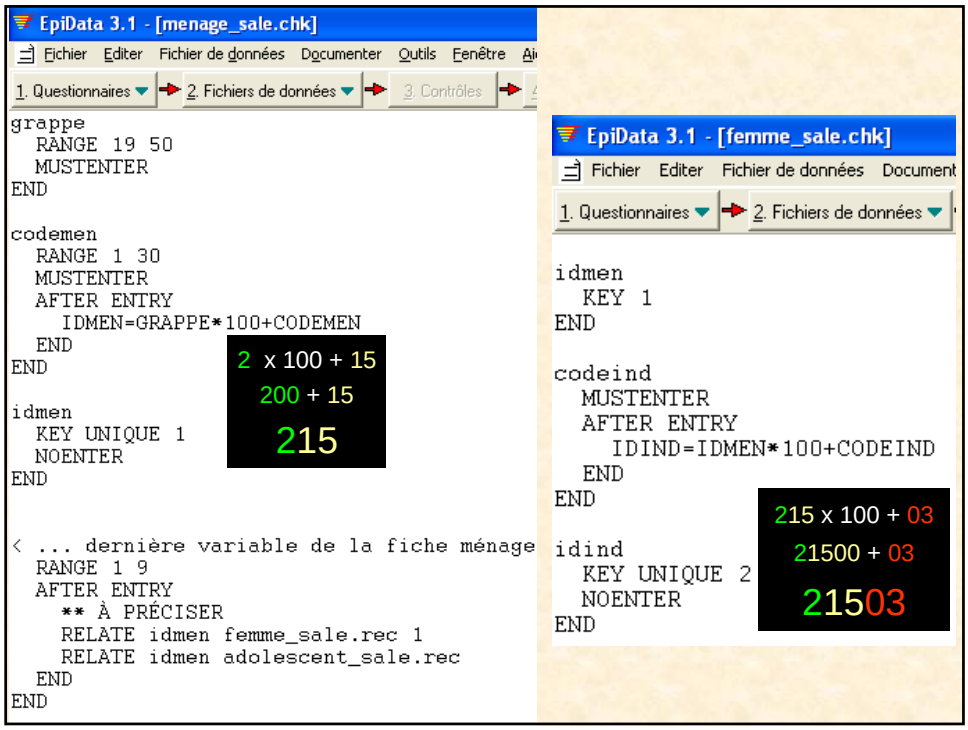


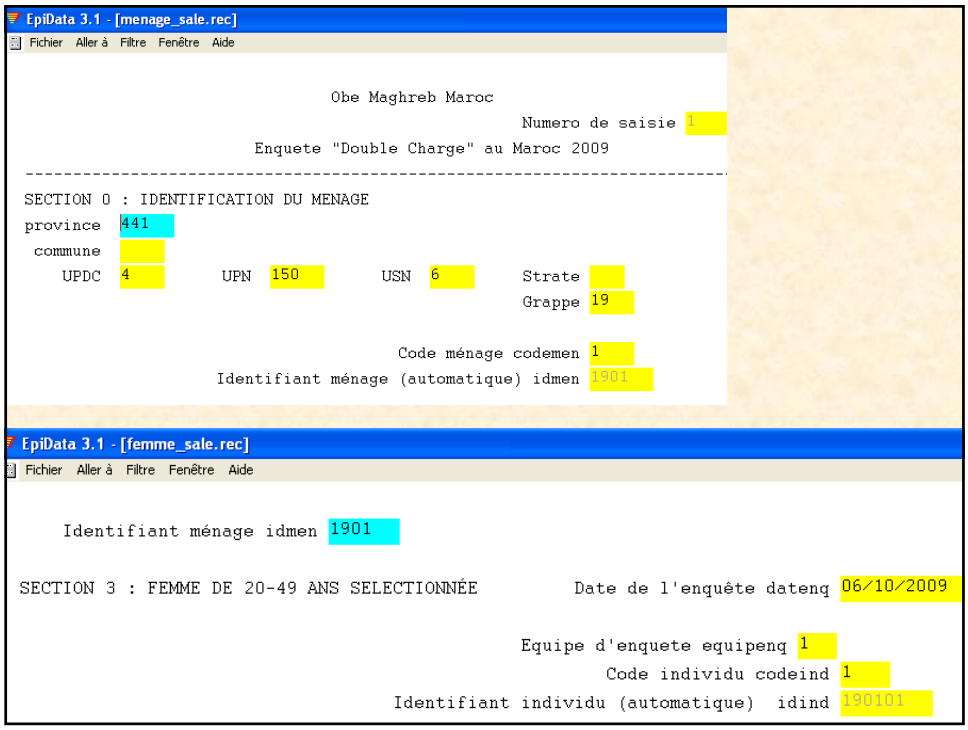
Modèle de données

- Identifiant ménage
  - grappe,
  - numéro de ménage dans la grappe

  exemple idmen = 215
- Identifiant femme
  - grappe,
  - numéro de ménage dans la grappe,
  - numéro de femme dans le ménage

  exemple idfem = 21503
- Prévoir les champs dans le masque


EpiData 3.1 - [menage_sale.chk]
Fichier Editer Fichier de données Documenter Outils Fenêtre
1. Questionnaires
2. Fichiers de données
3. Contrôles
grappe
RANGE 19 50
MUSTENTER
END
codemen
RANGE 1 30
MUSTENTER
AFTER ENTRY
IDMEN=GRAPPE*100+CODEMEN
END
END
idmen
KEY UNIQUE 1
NOENTER
END
2 x 100 + 15
200 + 15
215
< ... dernière variable de la fiche ménage
RANGE 1 9
AFTER ENTRY
** À PRÉCISER
RELATE idmen femme_sale.rec 1
RELATE idmen adolescent_sale.rec
END
END
EpiData 3.1 - [femme_sale.chk]
Fichier Editer Fichier de données Document
1. Questionnaires
2. Fichiers de données
idmen
KEY 1
END
codeind
MUSTENTER
AFTER ENTRY
IDIND=IDMEN*100+CODEIND
END
END
idind
KEY UNIQUE 2
NOENTER
END
215 x 100 + 03
21500 + 03
21503


EpiData 3.1 - [menage_sale.rec]
Fichier Aller à Filtre Fenêtre Aide
Obe Maghreb Maroc
Numero de saisie 1
Enquete "Double Charge" au Maroc 2009
SECTION 0 : IDENTIFICATION DU MENAGE
province 441
commune
UPDC 4
UPN 150
USN 6
Strate
Grappe 19
Code ménage codemen 1
Identifiant ménage (automatique) idmen 1901
EpiData 3.1 - [femme_sale.rec]
Fichier Aller à Filtre Fenêtre Aide
Identifiant ménage idmen 1901
SECTION 3 : FEMME DE 20-49 ANS SELECTIONNÉE
Date de l'enquête datenq 06/10/2009
Equipe d'enquete equipenq 1
Code individu codeind 1
Identifiant individu (automatique) idind 190101

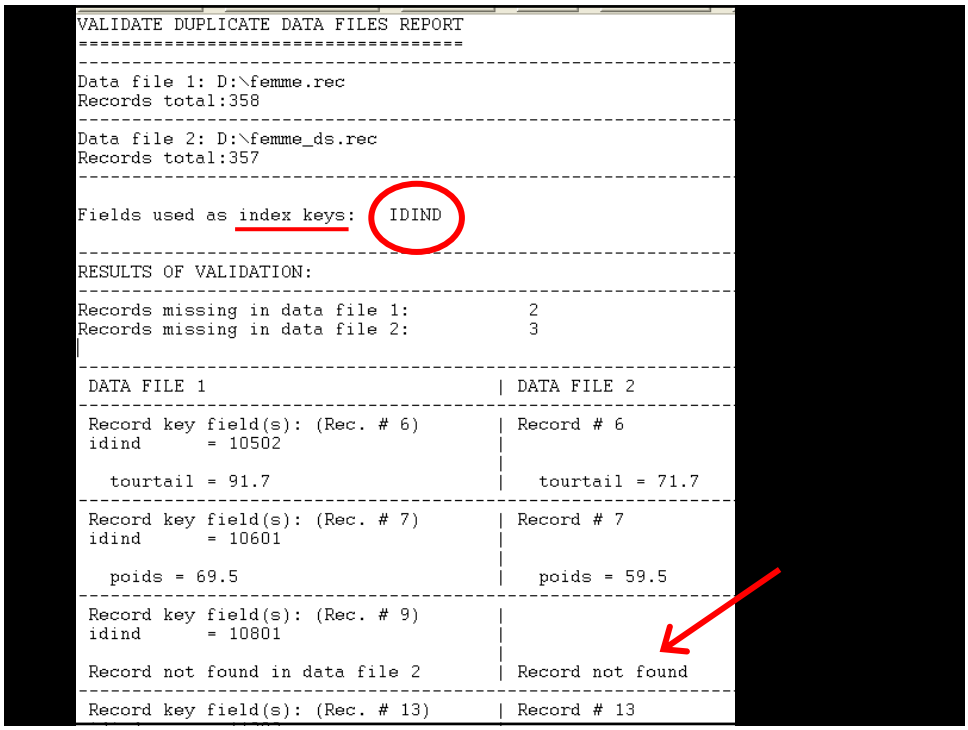
Détection et correction d'erreurs

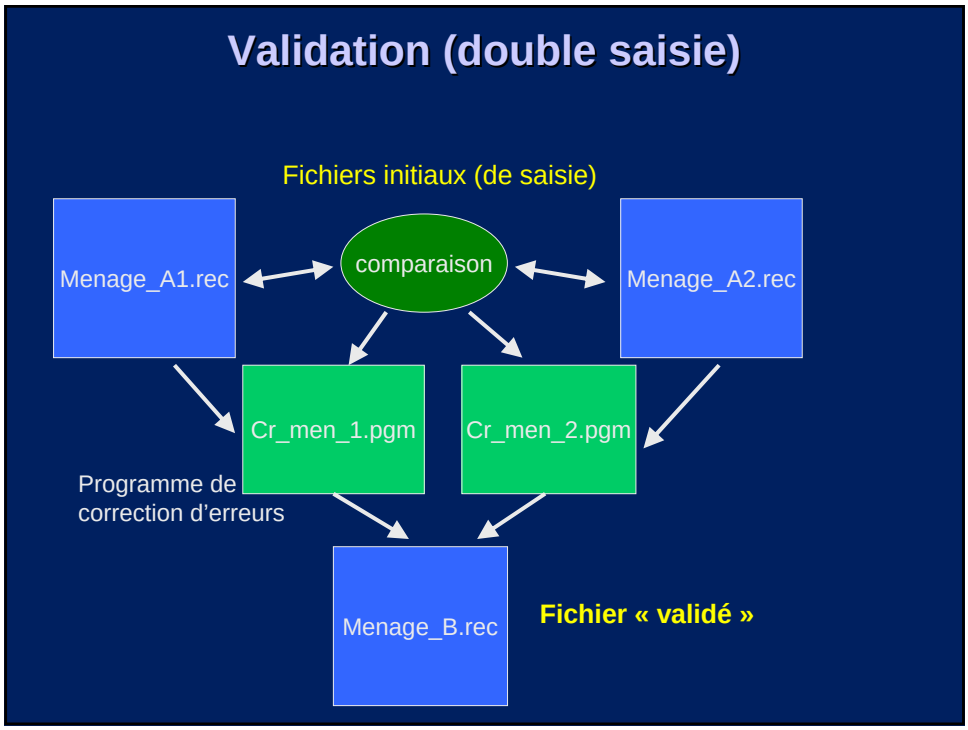
- **Détection & correction d'erreurs**
  pendant la saisie :
  contrôles d'entrée

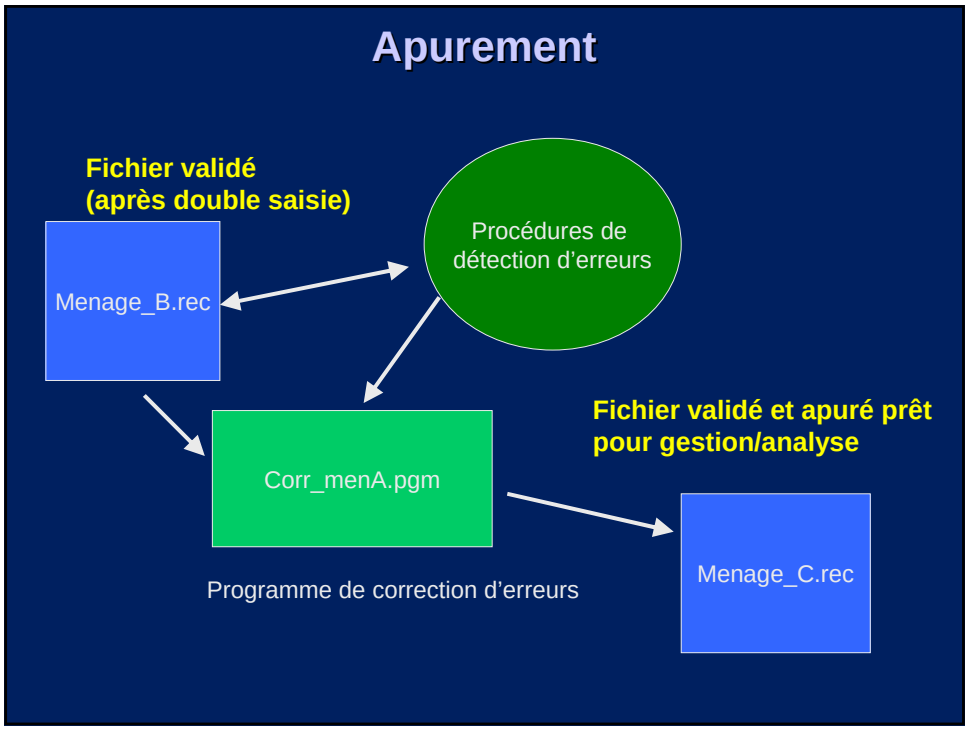
- **Détection & correction d'erreurs**
  après la saisie :
  validation (double saisie)
  apurement

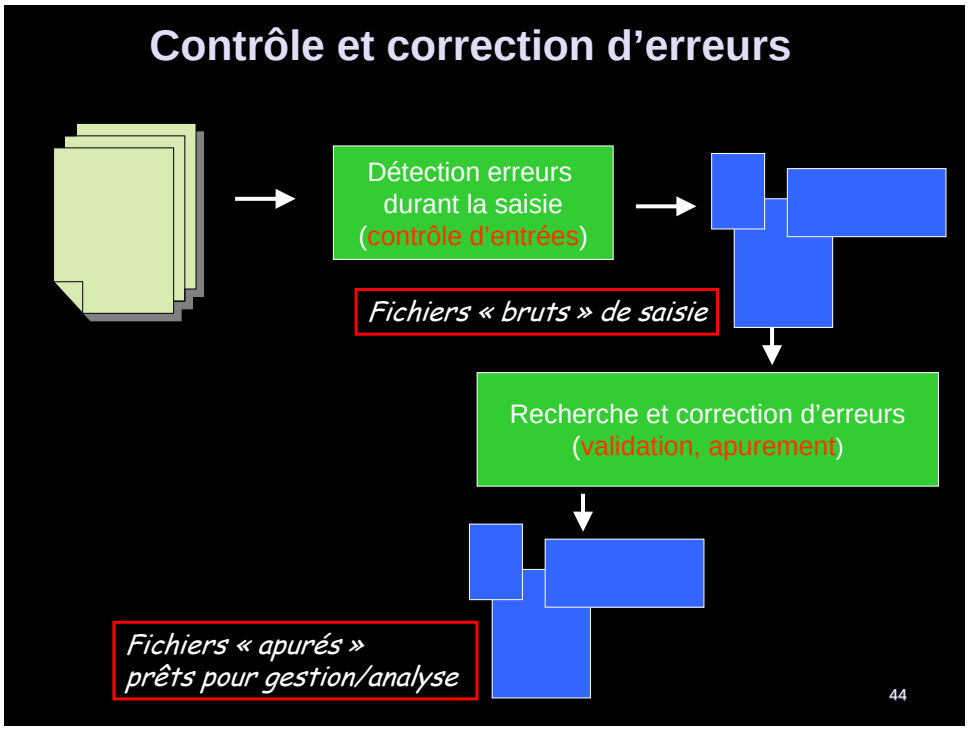
Validation

- Double saisie
- Comparaison des 2 fichiers .rec
- Quand différence, vérifier dans le questionnaire
- Correction
- re-validation ….etc…


VALIDATE DUPLICATE DATA FILES REPORT
Data file 1: D:\femme.rec
Records total:358
Data file 2: D:\femme_ds.rec
Records total:357
Fields used as index keys: IDIND
RESULTS OF VALIDATION:
Records missing in data file 1: 2
Records missing in data file 2: 3
DATA FILE 1
DATA FILE 2
Record key field(s): (Rec. # 6)
idind = 10502
tourtail = 91.7
Record # 6
tourtail = 71.7
Record key field(s): (Rec. # 7)
idind = 10601
poids = 69.5
Record # 7
poids = 59.5
Record key field(s): (Rec. # 9)
idind = 10801
Record not found in data file 2
Record not found
Record key field(s): (Rec. # 13)
Record # 13


Validation (double saisie)
Fichiers initiaux (de saisie)
Menage_A1.rec
comparaison
Menage_A2.rec
Cr_men_1.pgm
Cr_men_2.pgm
Programme de correction d'erreurs
Menage_B.rec
Fichier « validé »


Apurement
Fichier validé
(après double saisie)
Menage_B.rec
Procédures de
détection d'erreurs
Corr_menA.pgm
Programme de correction d'erreurs
Fichier validé et apuré prêt
pour gestion/analyse
Menage_C.rec


Contrôle et correction d'erreurs
Détection erreurs
durant la saisie
(contrôle d'entrées)
Fichiers « bruts » de saisie
Recherche et correction d'erreurs
(validation, apurement)
Fichiers « apurés »
prêts pour gestion/analyse
44