Projet Obe Maghreb

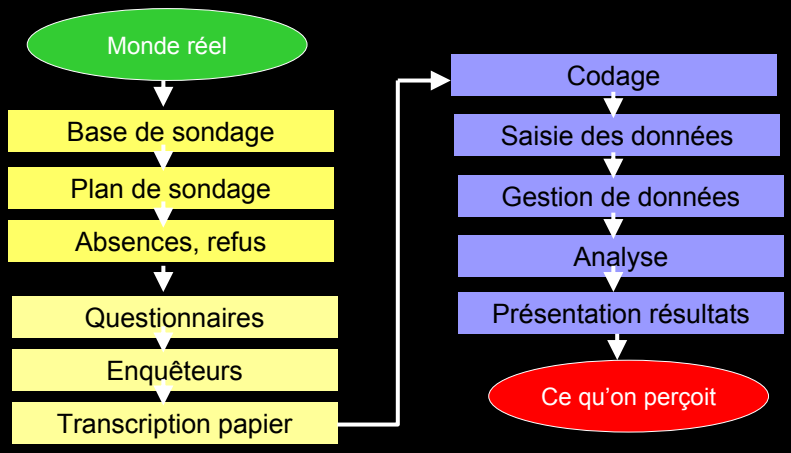
Ecole thématique gestion et analyse de données
20 au 29 avril 2010

Gestion et analyse de données d'enquêtes épidémiologiques

Modèle de données
Variables,unités statistiques, relations.

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

Modèle de données

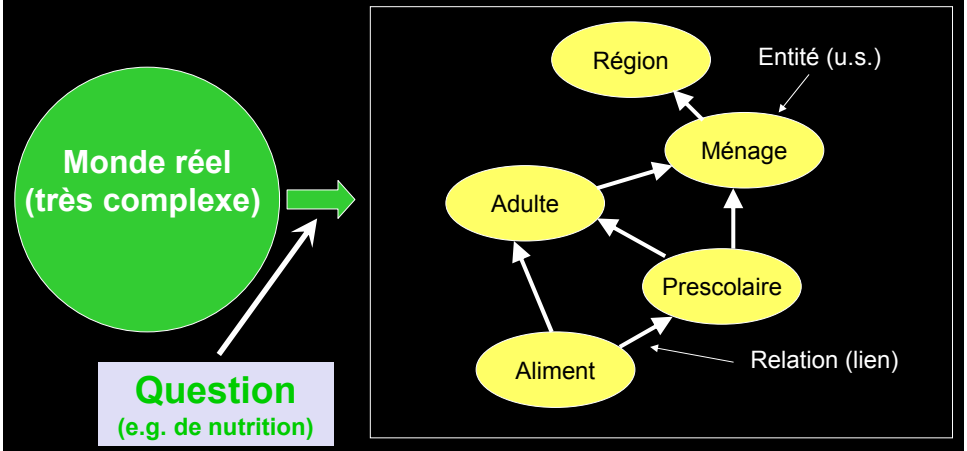


Modèle de données

- **Tout le processus** (même avant collecte des données)
  - échantillonnage (théorique, pratique)
  - **conception du questionnaire**
  - mise en pratique sur le terrain
  - **organisation de la saisie**
  - organisation de la base de données (interrogations possibles)
  - **analyses**
- **Eléments structurants communs**
  - **variable (attribut, caractère)**
  - unité statistique (observation, entité)
  - **modèle entités-relation (modèle de données)**
  - vocabulaires épidémiologique/statistique
    v.s. informatique (base de données relationnelle)

Modèle de données

- **Représentation schématique monde réel**



Modèle de données

- **Variable (attribut, caractère)**

Variables

- **Variables (attributs, caractères)**

  - caractéristiques d'intérêt pour l'étude (modèle causal, …)

  - poids, taille, âge, sexe, nombre de personnes du ménage, type d'habitat, CSP, dépense alimentaire mensuelle, distance du village au centre de santé, nombre de médecins /1000 h, date de naissance, région, nombre de supérettes, dépense énergétique journalière, attitude vis-à-vis de l'obésité fumeur (O/N),…

  - variables initiales : mesurées / observées / par interrogation
  items du questionnaire d'enquête
  e.g. sexe, poids, type d'habitation, revenus mensuels

  - variables dérivées : résultant d'un calcul ultérieur (indices, scores):
  e.g. IMC, dépense énergétique, attitude vis à vis obésité, indice taille-âge, score de niveau économique, qualité de vie

Variables

- **Différents types statistiques de variables**

- Quantitative . continue (taille en cm : 175,6 ; ingéré : 2456 kcal)
. discrête (nombre de personne dans ménage : 12)

Valeurs dans domaine (intervalle)

- Qualitative . ordonnée (état habitat : bon , moyen, mauvais)
. quelconque
(statut matrimonial : célibataire, marié, veuf, divorcé, autre)
. dichotomique (sexe : F/M, fumeur : oui/non)

Modalités exclusives, exhaustives (catégorie « autre »)

Modalités dans domaine (liste de modalités)

Variables

- **Différents types informatiques, dont :**

- numériques
  - . entiers : 0, 1, 15, 345
  - . réels : 3.56, 100, -2.5
  - . longueur, décimales
  - . opérations algébriques : +, x, -, /, log, …
  - . comparaisons : < >, <=, =, <>
- caractère
  - . contenu : obèse, oui, non, c2, 145, 4.5, …
  - . longueur
  - . pas d'opérations algébriques
  - . fonctions spécifiques (sous chaînes, …)

Variables

- **Notion de codage (ce qu'on va saisir)**

  - variable quantitative : unités
    taille : 1.756 m, 175.6 cm, 1756 mm

  - variable qualitative : codes (beaucoup de choix possibles)

  Satut matrimonial :

Valeurs	---------- Différents codages possibles ----------					
Célibataire	Célibataire	CEL	C	1	5	4
Marié(e)	Marié	MAR	M	2	4	5
Veuf(ve)	Veuf	VEU	V	3	3	6
Divorcé(e)	Divorcé	DIV	D	4	2	7
Autre	Autre	AUT	A	5	1	8

  - !!! codes ≠ valeurs de la variable (cf. analyse) !!!

  - documentation : unités, codes (variables de base ET calculées)

Modèle de données

- **Unité statistique (observation, entité)**

Unités statistiques

Unité statistique (u.s.)
(entité, observation, individu, enregistrement)

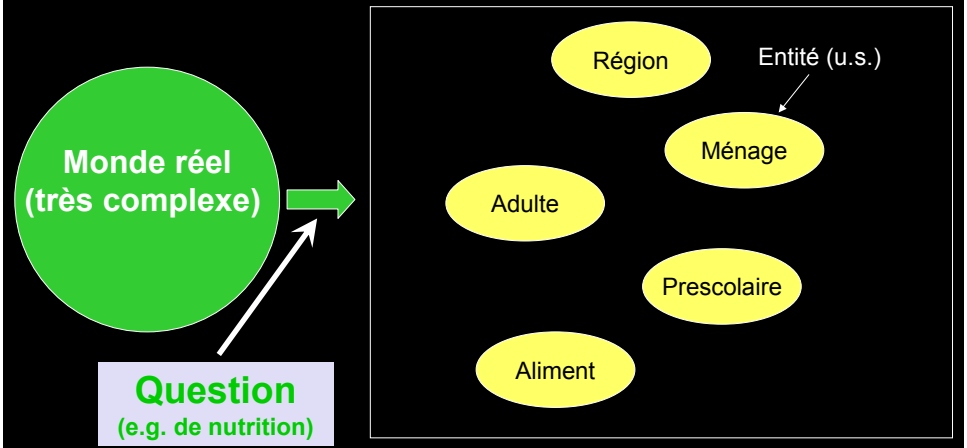
- La plus petite entité sur laquelle la valeur de la variable est définie
  (mesurée / observée / obtenue par interrogation / calculée)

- Dépend des variables
  . Revenu, type habitat, nbre de personnes : u.s. = un ménage
  . poids, taille, sexe : u.s. = une personne
  . nbre de médecins / 1000 h, PNB : u.s. = un pays
  . présence d'un centre de santé, nb épiceries : u.s. = une commune
  . énergie, lipides, glucides, protéines : u.s. = un aliment

- !!! Définition précises des u.s. (e.g. ménage) !!!

Unités statistiques

- Représentation schématique monde réel



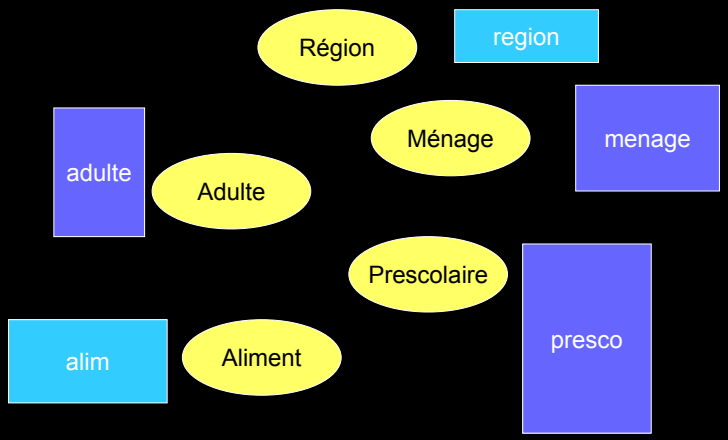
Unités statistiques enquêtées

- **Tableau u.s. x variables** (n=250 x p=10) : **adultes**

nusaiad	idadu	idmen	numad	datenq	sexe	age	statmat	gross	tailled
1	100010203	1000102	3	12/04/2000	1	45	2	2	165,4
2	100010206	1000102	6	12/04/2000	2	36	2	1	145,6
3	100020702	1000207	2	19/04/2000	1	34	1	2	170,2
4	100020703	1000207	3	19/04/2000	1	52			
5	100020704	1000207	4	20/04/2000	2	65	3	1	159,8
6	100031706	1000317	6	04/04/2000	2	38	4	2	
7	100031803	1000318	3	05/04/2000	2	37	2	2	174,6
8	100031806	1000318	6	05/04/2000	1	44	1	2	169,0
…	…	…	…	…	…		…	…	…

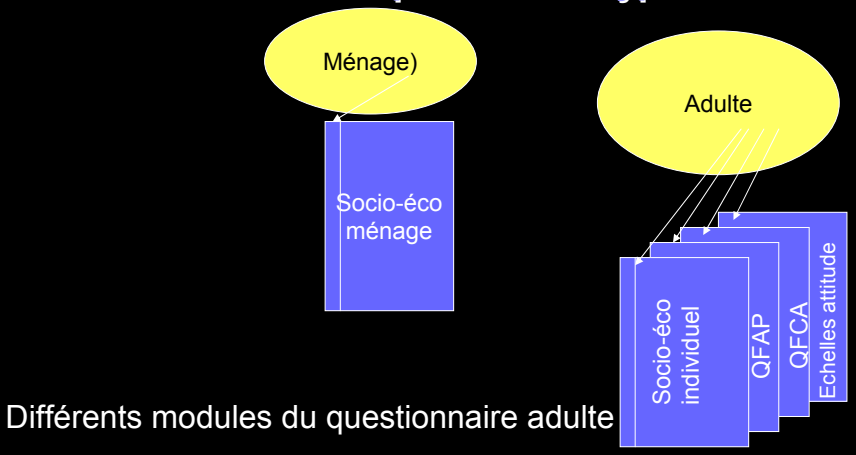
Unités statistiques

- **Chaque niveau d'u.s. => table u.s. x variables**



Unités statistiques

- **Différentes tables pour même type d'u.s.**



Différents modules du questionnaire adulte

Unités statistiques

- **Notation**

  adultes(nusaiad,idadu, idmen, numad, datenq, sexe, age, statmat, gross, tailled)

- **Informatique (EpiInfo, EpiData, SAS, Stata,…)**

  - tableau de données => fichier (exception : MS Access)

  - format interne / nom : dépend du logiciel
    - EpiInfo/EpiData : extension .rec (adultes.rec)
    - SAS : extension .ssd (adultes.ssd)
    - SPSS : extension .sav (adultes.sav)
  - nom du fichier / des colonnes :
    - contraintes particulières suivant logiciel

- **Documentation :**

  dictionnaire de variables

Dictionnaire de variables

Ordre dans le fichier	nom	Contenu	format	Unités / codes	n	Pgm de création
1	nusaiad	Numéro séquentiel de saisie adulte	Numérique 4.	Sans objet	250	adultes.qes
2	idadu	Identifiant unique d'adulte	Numérique 4.	Sans objet	250	adultes.qes adultes.chk
3	idmen	Identifiant de ménage	Numérique 4.	Sans objet	250	adultes.qes adultes.chk
4	numad	Numéro adulte dans ménage	Numérique 2.	Sans objet	250	adultes.qes
5	datenq	Date d'enquête	Date (dmy 10)	jj/mm/aaaa	250	adultes.qes
6	sexe	Sexe de la personne	Numérique 1.	1: masculin 2: féminin	250	adultes.qes
7	age	Age de la personne	Numérique 3.	années révolues	245	adultes.qes
8	statmat	Statut matrimonial	Numérique 1.	1: célibataire 2: marié 3: …..	240	adultes.qes
9	gross	Grossesse visible	Numérique 1.	1:oui 2:non	145	adultes.qes
10	tailled	Taille debout	Numérique 5.1	cm	238	adultes.qes

Unités statistiques enquêtées

Tableau u.s. x variables (n=250 x p=10) : adultes

nusaiad	idadu	idmen	numad	datenq	sexe	age	statmat	gross	tailled
1	100010203	1000102	3	12/04/2000	1	45	2	2	165,4
2	100010206	1000102	6	12/04/2000	2	36	2	1	145,6
3	100020702	1000207	2	19/04/2000	1	34	1	2	170,2
4	100020703	1000207	3	19/04/2000	1	52			
5	100020704	1000207	4	20/04/2000	2	65	3	1	159,8
6	100031706	1000317	6	04/04/2000	2	38	4	2	
7	100031803	1000318	3	05/04/2000	2	37	2	2	174,6
8	100031806	1000318	6	05/04/2000	1	44	1	2	169,0
…	…	…	…	…	…		…	…	…

Unités statistiques enquêtées

- **Tableau u.s. x variables** (n=250 x p=10) : **adultes**

nusaiad	idadu	idmen	numad	datenq	sexe	age	statmat	gross	tailled
1	100010203	1000102	3	12/04/2000	1	45	2	2	165,4
2	100010206	1000102	6	12/04/2000	2	36	2	1	145,6
3	100020702	1000207	2	19/04/2000	1	34	1	2	170,2
4	100020703	1000207	3	19/04/2000	1	52			
5	100020704	1000207	4	20/04/2000	2	65	3	1	159,8
6	100031706	1000317	6	04/04/2000	2	38	4	2	
7	100031803	1000318	3	05/04/2000	2	37	2	2	174,6
8	100031806	1000318	6	05/04/2000	1	44	1	2	169,0
…	…	…	…	…	…		…	…	…

Unités statistiques

- **Identifiant (clé primaire, key)**

- variable ayant une valeur distincte pour chaque u.s. du tableau

- parfois simple numéro (1 à n)

- souvent combinaison de plusieurs variables – e.g. Tunisie
  région, gouvernorat, district, commune, ménage, personne
  02 15 10 09 267 05
  => valeur de l'identifiant 0215100926705

- Notation symbolique « souligné » : idadu

- !!! existence et unicité de la clé dans chaque tableau !!!
  (intégrité d'entité)

Unités statistiques enquêtées

- **Tableau u.s. x variables** (n=250 x p=10) : **adultes**

nusaiad	idadu	idmen	numad	datenq	sexe	age	statmat	gross	tailled
1	100010203	1000102	3	12/04/2000	1	45	2	2	165,4
2	100010206	1000102	6	12/04/2000	2	36	2	1	145,6
3	100020702	1000207	2	19/04/2000	1	34	1	2	170,2
4	100020703	1000207	3	19/04/2000	1	52			
5	100020704	1000207	4	20/04/2000	2	65	3	1	159,8
6	100031706	1000317	6	04/04/2000	2	38	4	2	
7	100031803	1000318	3	05/04/2000	2	37	2	2	174,6
8	100031806	1000318	6	05/04/2000	1	44	1	2	169,0
…	…	…	…	…	…		…	…	…

Unités statistiques

- **Existence et unicité de la clé**

  - Conception du questionnaire

    prévoir et documenter la construction des identifiants

  - Conception des utilitaires de saisie

    . attention spécifique aux identifiants

    . sécurité : numéro de saisie

  - Validation, apurement : vérifier existence et unicité

  - Gestion de données : ne pas oublier identifiant(s) « en route »

Modèle de données

)

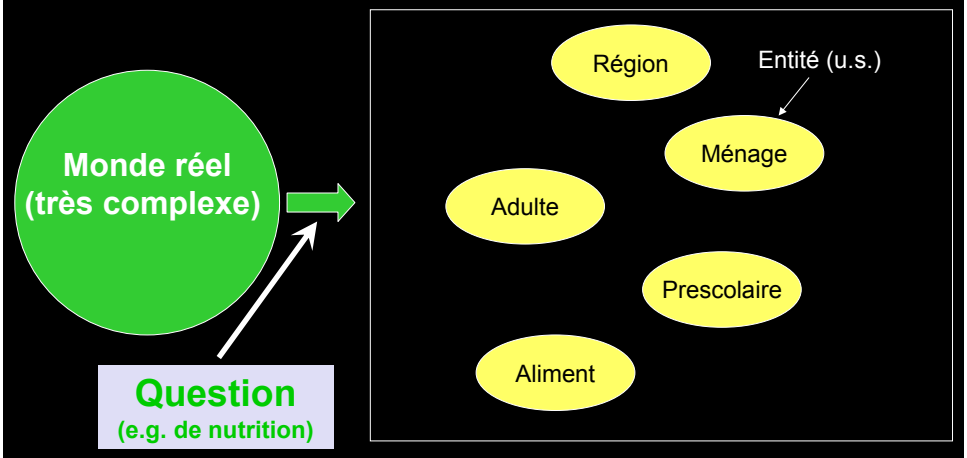
- Modèle entités-relation (modèle de données)

Modèle entités-relations

- Unité statistique (u.s.)
  - Dépend des variables
    - . Revenu, type habitat, nbre de personnes : u.s. = un ménage
    - . poids, taille, sexe : u.s. = une personne
    - . nbre de médecins / 1000 h, PNB : u.s. = un pays
    - . présence d'un cente de santé, nb épicéries : u.s. = une commune
    - . énergie, lipides, glucides, protéines : u.s. = un aliment
- Dans une étude :
  - plusieurs types d'u.s. : e.g. ménage, adulte, prescolaire, aliment
  - relations entre les u.s. :
    - . adulte est dans ménage
    - . prescolaire est dans ménage
    - . prescolaire est enfant de adulte
    - . adulte a consommé aliment
- u.s. et relations : modèle entités-relations
- Modèle relationnel, bases de données relationnelles

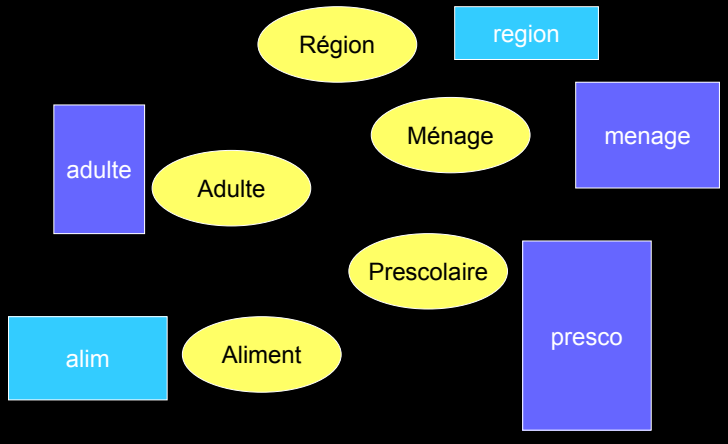
Modèle entités-relations

- Représentation schématique monde réel



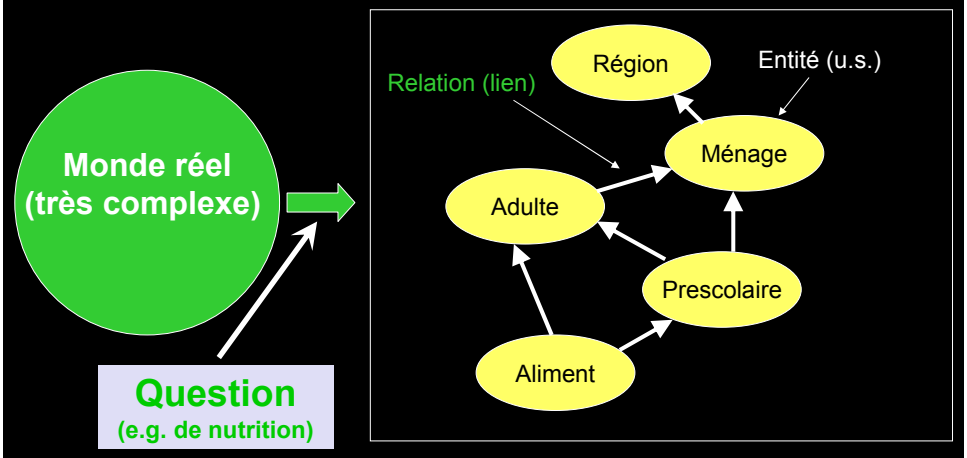
Modèle entités-relations

- Chaque niveau d'u.s. => table u.s. x variables



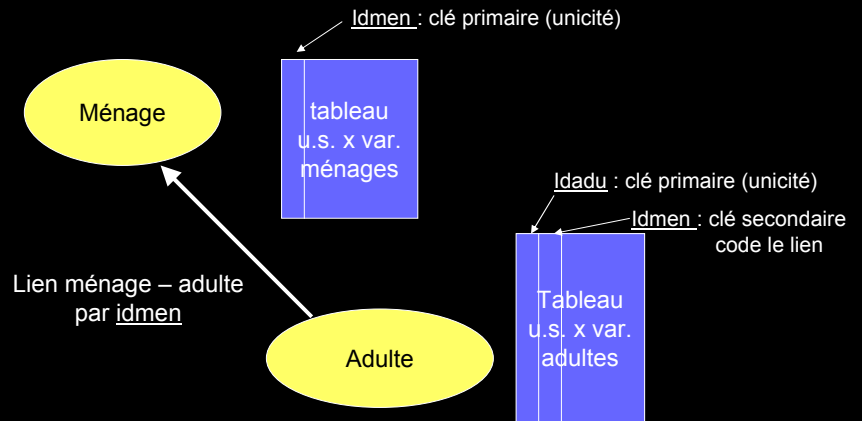
Modèle entités-relations

- **Représentation schématique monde réel**



Modèle entités-relations

- **Lien « adulte appartient à ménage »**



Modèle entités-relations

- **Relation**

  Exemple : «adulte appartient à ménage»

  - adulte (idadu, idmen, datenq, sexe, age, statmat, gross, tailled)

idadu	idmen	datenq	sexe	age	statmat	...
100010203	1000102					
100010206	1000102					
100031706	1000317					
100031803	1000318					

- idadu : identifiant / clé primaire (unicité dans adulte)
- idmen : identifiant / clé secondaire (non unicité dans adulte, unicité dans menage)

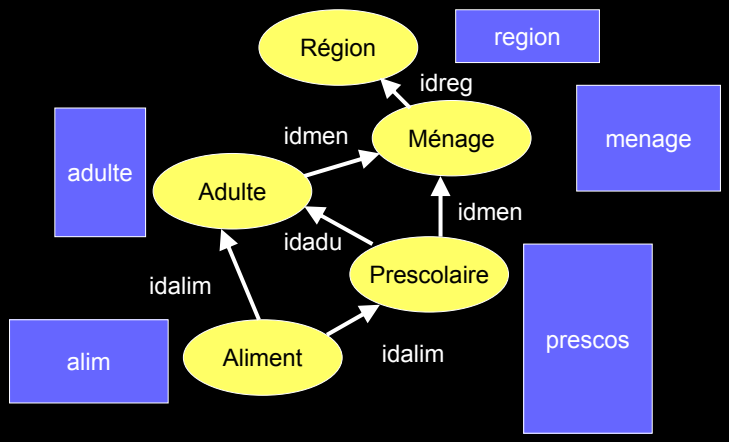
Unités statistiques

- **Tableau u.s. x variables** (n=250 x p=10) : **adultes**

nusaiad	idadu	idmen	numad	datenq	sexe	age	statmat	gross	tailled
1	100010203	1000102	3	12/04/2000	1	45	2	2	165,4
2	100010206	1000102	6	12/04/2000	2	36	2	1	145,6
3	100020702	1000207	2	19/04/2000	1	34	1	2	170,2
4	100020703	1000207	3	19/04/2000	1	52			
5	100020704	1000207	4	20/04/2000	2	65	3	1	159,8
6	100031706	1000317	6	04/04/2000	2	38	4	2	
7	100031803	1000318	3	05/04/2000	2	37	2	2	174,6
8	100031806	1000318	6	05/04/2000	1	44	1	2	169,0
...	...	...	...	...	...		...	...	...

Modèle entités-relations

- **Explicitation des liens**



Modèle entités-relations

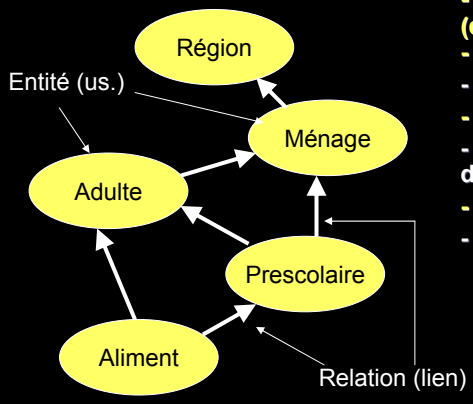
- **u.s. et liens**
  - adulte (idadu, idmen, datenq, agea, sexea, …)
    ou
    adulte (idadu, idmen, idreg, datenq, agea, sexea, …)

  - presco (idenf, idadu, datenq, datnai, …)
    ou
    presco (idenf, idmen, idreg, idadu, datenq, datnai, …)

- **Cas fréquent**
  - idmen contient idreg : 1000102
  - idadu contient idreg, idmen : 100010206
  - idenf contient idreg, idmen, idadu : 10001020602

Modèle entités-relations

- **Schéma**



- **Sous-jacent à :**

- la conception du questionnaire (différents modules)
- l'échantillonnage
- les activités de terrain
- l'organisation de la saisie
- l'organisation de la base de données (interrogations possibles)
- analyses
- présentation des résultats

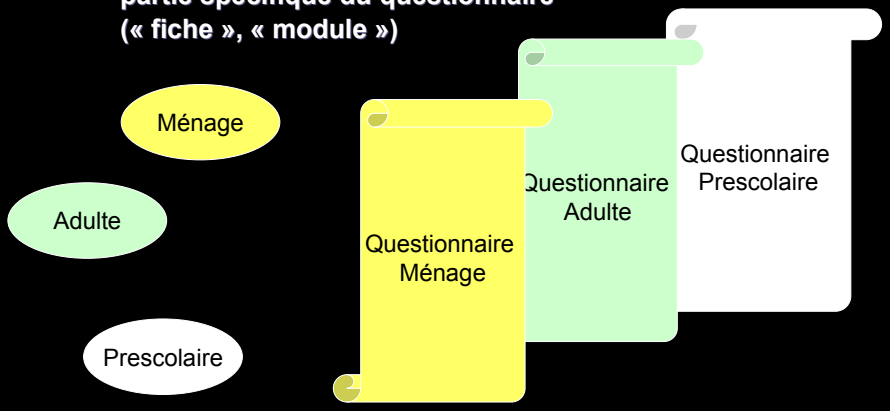
!!! Identifiants !!!

Modèle de données

- **Conception du questionnaire**

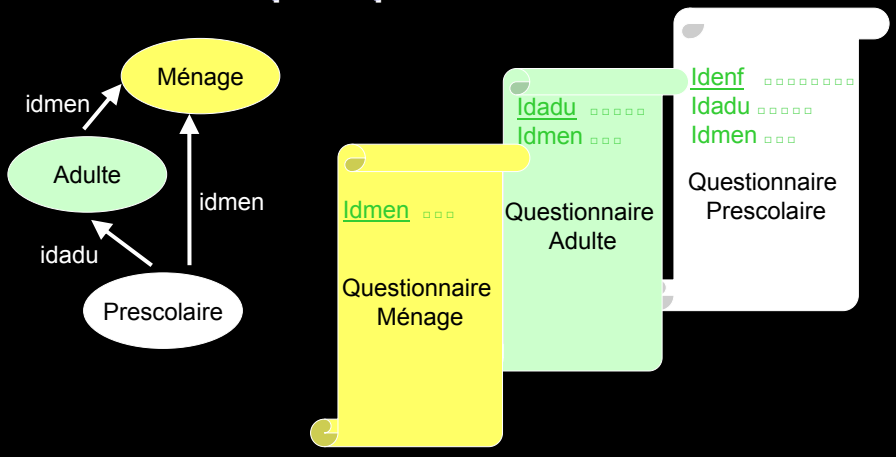
Conception du questionnaire
Unités statistiques

- **Chaque type d'unité statistique :**
  partie spécifique du questionnaire
  (« fiche », « module »)

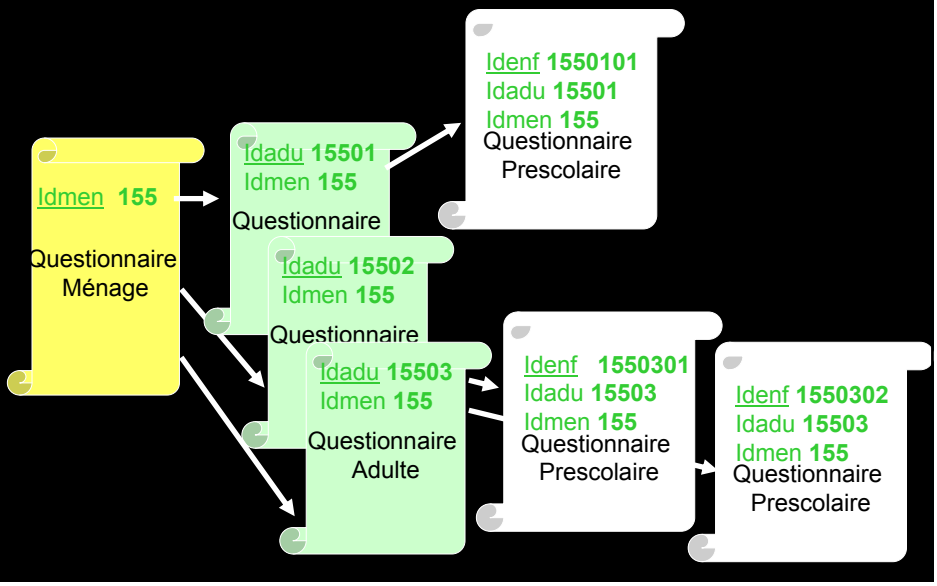


Conception du questionnaire
Relations

- **Identifiants principaux et secondaires**

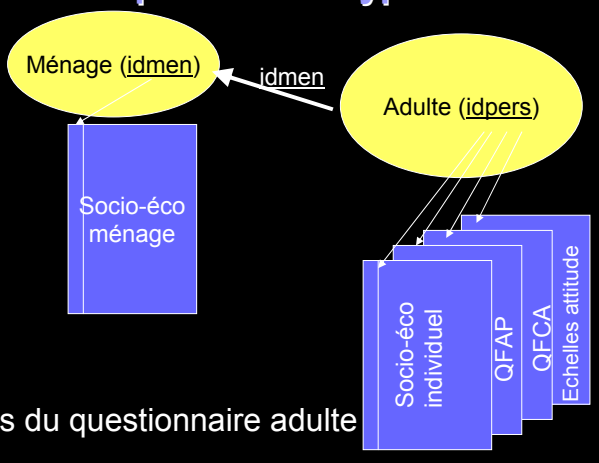


Conception du questionnaire
Relations



Conception du questionnaire
Relations

- **Différentes tables pour même type d'u.s.**



Différents modules du questionnaire adulte

Conception du questionnaire
Variables

- **Choix des variables d'intérêt**
  - objectifs de l'enquête
    - . demande d'information
    - . modèle causal / conceptuel
    - . bibliographie
  - type d'unité statistique
  - plan d'analyse
- **Distinguer variables initiales / dérivées**
- **Choix des modalités et codages**
  - variable quantitatives : unités
  - variables qualitatives : modalités, codages
  - questions ouvertes / fermées
  - codage des données manquantes

Conception du questionnaire
Variables

- **Démarche « ingénierie inverse »**
  1. information souhaitée / question posée ?
  2. tableau, analyse, graphique nécessaire ?
  3. variables dérivées nécessaires ?
  4. variables à recueillir / mesurer ?
     items du questionnaire
  5. codages ?
- **Exemple**

Conception du questionnaire
Variables

1- Information souhaitée ?

« Effet du niveau économique du ménage sur le retard de croissance en taille des enfants prescolaires »

(e.g. Maroc milieu urbain)

Conception du questionnaire
Variables

2. Analyse, type de tableau à construire ?

Retard de taille prescolaires

		Prévalence T.A. <-2 Z
bas	(n=900)	30% (e.g.)
moyen	(n=900)	20% (e.g.)
élevé	(n=900)	5% (e.g.)

Terciles de niveau économique ménage

Conception du questionnaire
Variables

3- Variables dérivées nécessaires ?

- retard de taille : indice taille pour âge <-2
  => indice taille pour âge à calculer

- niveau économique bas /moyen/élevé :
  terciles d'un indice de niveau économique du ménage
  => cet indice à construire :
  réfléchir comment caractériser le niveau économique du ménage en fonction du contexte

(e.g. nombre de biens possédés)

Conception du questionnaire
Variables

4- Variables à recueillir par questionnaire ?

- fiche enfant prescolaire
  sexe, taille,
  date de naissance, date d'enquête (âge)
- fiche ménage
  lave linge, réfrigérateur, TV, parabole, ordinateur, voiture….

! contexte

Conception du questionnaire
Variables

5- Codages ?

- fiche enfant prescolaire

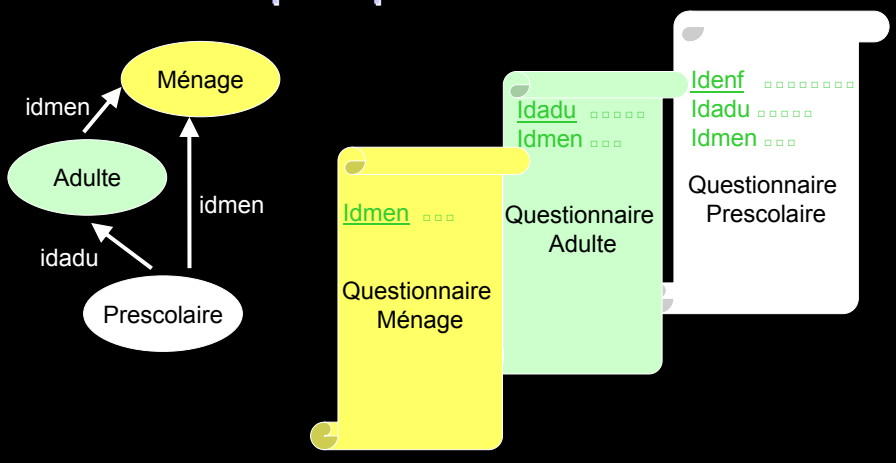
sexe (1:masculin, 2:féminin)	/_/
taille (en cm)	/_/_/_/_/_/ . /_/
date de naissance (jjmmaaaa)	/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/
date d'enquête (jjmmaaaa)	/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/

- fiche ménage

lave-linge	(1:oui/2:non,9:nsp)	/_/
réfrigérateur	(1:oui/2:non,9:nsp)	/_/
TV	(1:oui/2:non,9:nsp)	/_/
parabole	(1:oui/2:non,9:nsp)	/_/
…		

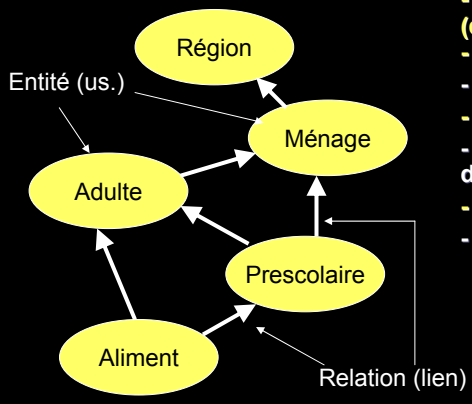
Conception du questionnaire
Relations

- **Identifiants principaux et secondaires**



Modèle de données

- **Schéma**



- **Sous-jacent à :**
  - la conception du questionnaire (différents modules)
  - l'échantillonnage
  - les activités de terrain
  - l'organisation de la saisie
  - l'organisation de la base de données (interrogations possibles)
  - analyses
  - présentation des résultats

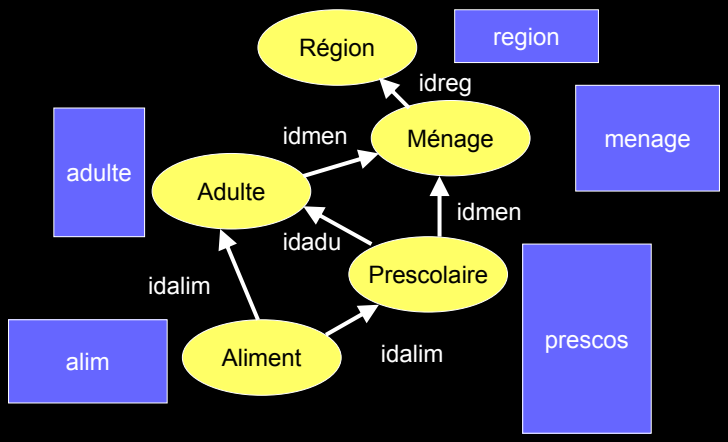
!!! Identifiants !!!

Modèle de données

- **Indispensable d'expliciter le modèle dès le début de l'étude (protocole)**
  - Identification précise des u.s. (e.g. ménage) et variables
  - Relations nécessaires pour l'étude
- **Identifiants**
- **Documentation de la base**
  - définition précise des u.s. (entités)
  - schéma du modèle de données (tables, relations)
  - dictionnaires de variables (nom, contenu, type, codes, …)

Modèle de données

- **Base de données relationelle**



Modèle de données

- **Règles d'intégrité**

1. Intégrité d'entité : existence et unicité de la clé (identifiant)

2. Intégrité de domaine : valeurs des variables dans domaines

3. Intégrité de référence :

adulte(idadu, idmen, sexe, age, …)
menage(idmen,nbpers, habitat, …)

Problème de référence si :

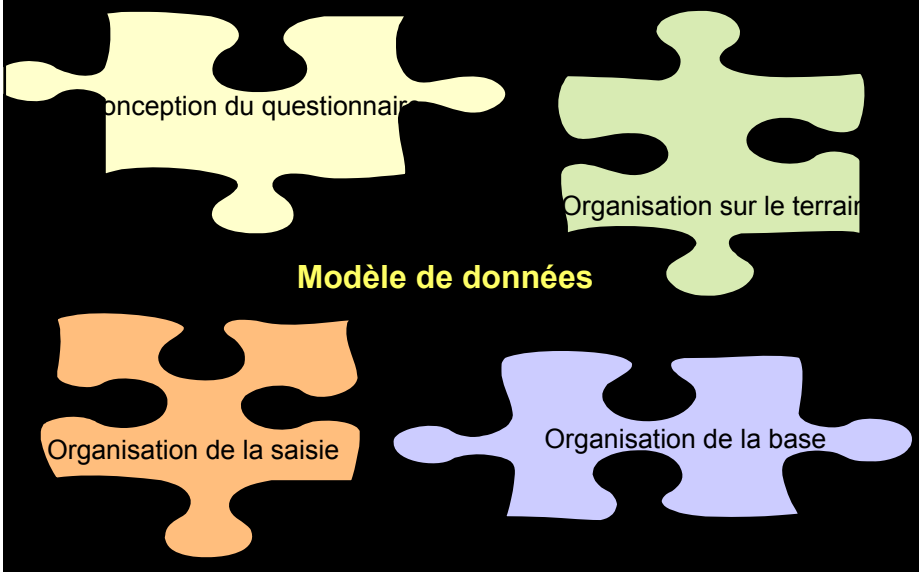
valeur de idmen . présente dans adulte
. absente dans menage

Maintenir cohérence (suppression)

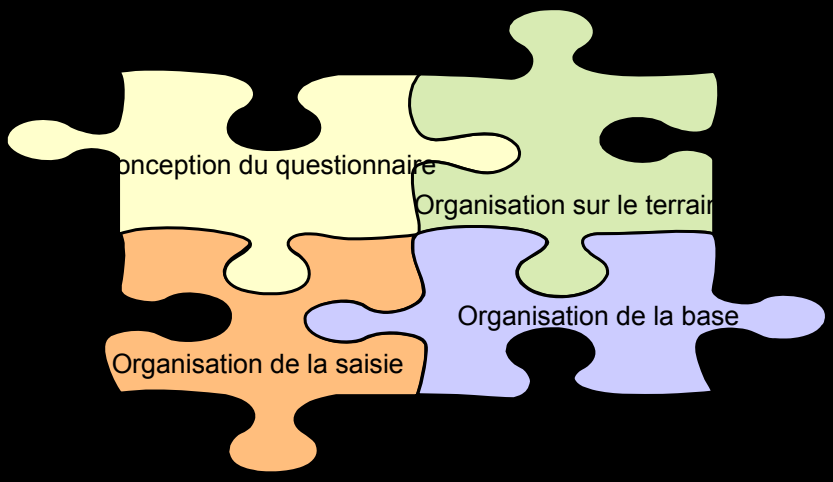
Modèle de données

- **Utilisation pratique de la base de données**
- **Préparation des données avant analyse**
  - sélections (invidus, variables)
  - mise en relation, fusion de fichiers
  - calcul de nouvelles variables
  - recodages
  - documentation
- **Langage de manipulation /gestion de données**
- **Outil logiciel** (seulement un outil *vs* concepts !)
  - SGBD (e.g. MS- Access, Oracle, …)
  - gestion de données dans logiciels statistiques (SAS, Stata, SPSS)
    ou généralistes (EpiDataAnalysis)

Modèle de données



Modèle de données



Modèle de données

Fin