DES BASES DE DONNEES

POUR LES ETUDES STATISTIQUES

Jean-François REY, Kathy CHAPELAIN, Jacques VAILLÉ

ISAB - Institut Supérieur Agricole de Beauvais
Mél : jfrancois.rey@isab.fr

Résumé : Cet article présente un système général pour l'expérimentation, avec base de données et traitements associés, ainsi qu'une réflexion et des réalisations sur l'organisation des données d'enquêtes et de panels incluant le problème des questions hiérarchisées.

Ces réalisations reposent sur une démarche où les méthodes de conception des systèmes d'information et des bases de données sont intégrées dans la méthodologie des études statistiques. On présente et justifie les principes qui sous-tendent ces méthodes et on discute le choix des outils, logiciel spécifique ou progiciel, tableur ou SGBD.

Mots-clés : *système d'information, base de données, paramétrage, SGBD, statistique, expérimentation, enquête, panel, questions hiérarchisées*

TABLE DES MATIERES

1 Introduction 73
2 Les principes 74
2.1 La modélisation des concepts et des données 75
2.2 La structuration des données en une base de données 77
2.3 La généralisation du problème et le paramétrage 80
3 Justification et avantages de l'approche préconisée 80
3.1 La structuration de la base de données en deux étapes 80
3.2 Un niveau de généralité adéquat 81
3.3 Les données ne sont pas a priori rassemblées en un seul tableau ! 81
3.4 Intérêt du recours à un SGBD 84
4 Modèles de données pour les questionnaires 85
4.1 L'enquête occasionnelle 85
4.2 Le panel 89
5 Le problème des questions hiérarchisées 92
5.1 La "structure réflexive" 92
5.2 L'emploi d'un code décimal hiérarchisé 93
6 Applications 95
6.1 L'expérimentation 96
6.2 Les enquêtes et panels 99
6.3 Les questions hiérarchisées dans un système de type administratif 100
7 Pour conclure 102
7.1 Méthodologie et transdisciplinarité 102
7.2 Modèle ou méta-modèle ? mais organisation découlant du problème réel ! 102
7.3 Logiciel spécifique ou progiciel ? SGBD ou tableur ? 103
8 Références 104

1. Introduction

Le traitement statistique des données fait appel aux méthodes et outils de la statistique et de l'informatique. Des rapprochements entre le vocabulaire du statisticien et celui de l'informaticien sont faciles à réaliser lorsqu'il s'agit de la gestion des données à traiter : pour le statisticien, un tableau décrit les informations, sous forme de variables mesurées, d'un ensemble d'individus ; pour l'informaticien, une table contient des enregistrements, eux-mêmes composés de divers champs (ou rubriques) renseignés par des valeurs.
Et si, entre eux, les missions et compétences diffèrent pour atteindre l'objectif visé lors d'une étude statistique, une coordination est cependant indispensable pour le bon aboutissement de l'étude.

Disciplines destinées par essence à être appliquées, la statistique et l'informatique doivent rester toutes deux à proximité des domaines et problèmes qu'elles servent à étudier. Et, pour mener des investigations ou des traitements parfois complexes, le recours aux méthodes leur est indispensable. Aussi, mettent elles en œuvre à la fois les démarches, les langages et les outils propres d'une part à chacune d'elles et d'autre part aux divers champs de connaissance ou d'activité concernés. Mais, la "véritable" méthode n'est-elle pas celle qui permet de progresser de façon transdisciplinaire dans le savoir ou la maîtrise relatifs au domaine abordé ?

Cet article voudrait contribuer au rapprochement des diverses approches en trois temps :
D'abord, en plaidant pour un usage raisonné des méthodes des systèmes d'information. Nous montrerons qu'elles s'inscrivent pleinement dans la méthodologie générale d'approche des problèmes et qu'elles contribuent à la qualité. Les principes que nous avons choisi de présenter (ou de rappeler) feront l'objet du chapitre 2 ; ils seront justifiés au chapitre 3.
Puis, nous présenterons des structures de données possibles pour gérer les questionnaires (enquêtes ou panels) au chapitre 4 et, au chapitre 5, deux approches pour la représentation et le traitement des questions hiérarchisées.
Enfin, en illustration, nous retraduirons l'expérience et quelques réalisations d'une équipe d'informaticiens et statisticiens qui ont eu à réunir leurs compétences pour la réalisation d'études.
Cette équipe est celle des enseignants-consultants et ingénieurs de la cellule STID (Statistique et Traitement Informatique des Données) de l'ISAB (Institut Supérieur Agricole de Beauvais).
Ses actions sont :

- le conseil et la réalisation d'études (une vingtaine d'études ou de projets conduits ou réalisés par an, sans compter les conseils ponctuels fournis aux collègues enseignants-chercheurs ou aux étudiants),
- la formation initiale des ingénieurs, avec une pédagogie reposant sur cette expérience de conseil [GRE98] (avec, actuellement, la participation de l'équipe aux travaux des Cercles d'Excel'ense [CERCL]),
- et la formation continue, notamment avec le concept de sessions de formation-conseil où les problèmes apportés par les stagiaires sont traités au moyen des outils méthodologiques introduits en première partie de la session. Ces formations sont destinées tantôt à un public de chargés d'études technico-économiques, tantôt à des chercheurs et thésards en sciences de la vie.

Les applications réalisées seront présentées au chapitre 6 ; au nombre de trois, elles concernent deux thèmes principaux : celui de l'expérimentation scientifique et celui de la gestion de questionnaires soumis à des évolutions au cours du temps.

Enfin, pour conclure, une discussion mettra en évidence les questions que peut se poser le statisticien : les choix à effectuer, l'évaluation des potentialités et risques, les conditions requises.

2. Les principes

Dans ce chapitre, nous présentons un ensemble de principes, généralement mis en œuvre dans les systèmes d'information, qui président à la conduite d'une étude quantitative dans les phases situées en amont des traitements statistiques.
Ces principes sont énoncés en trois parties : la modélisation des concepts et des données, la structuration de ces données en base de données, et le caractère plus ou moins général et paramétrable que l'on peut souhaiter pour cette organisation. Ils sont aussi reliés les uns aux autres en référence aux étapes de la méthodologie générale de résolution des problèmes dont nous proposons une représentation du cheminement sur la figure 1 :

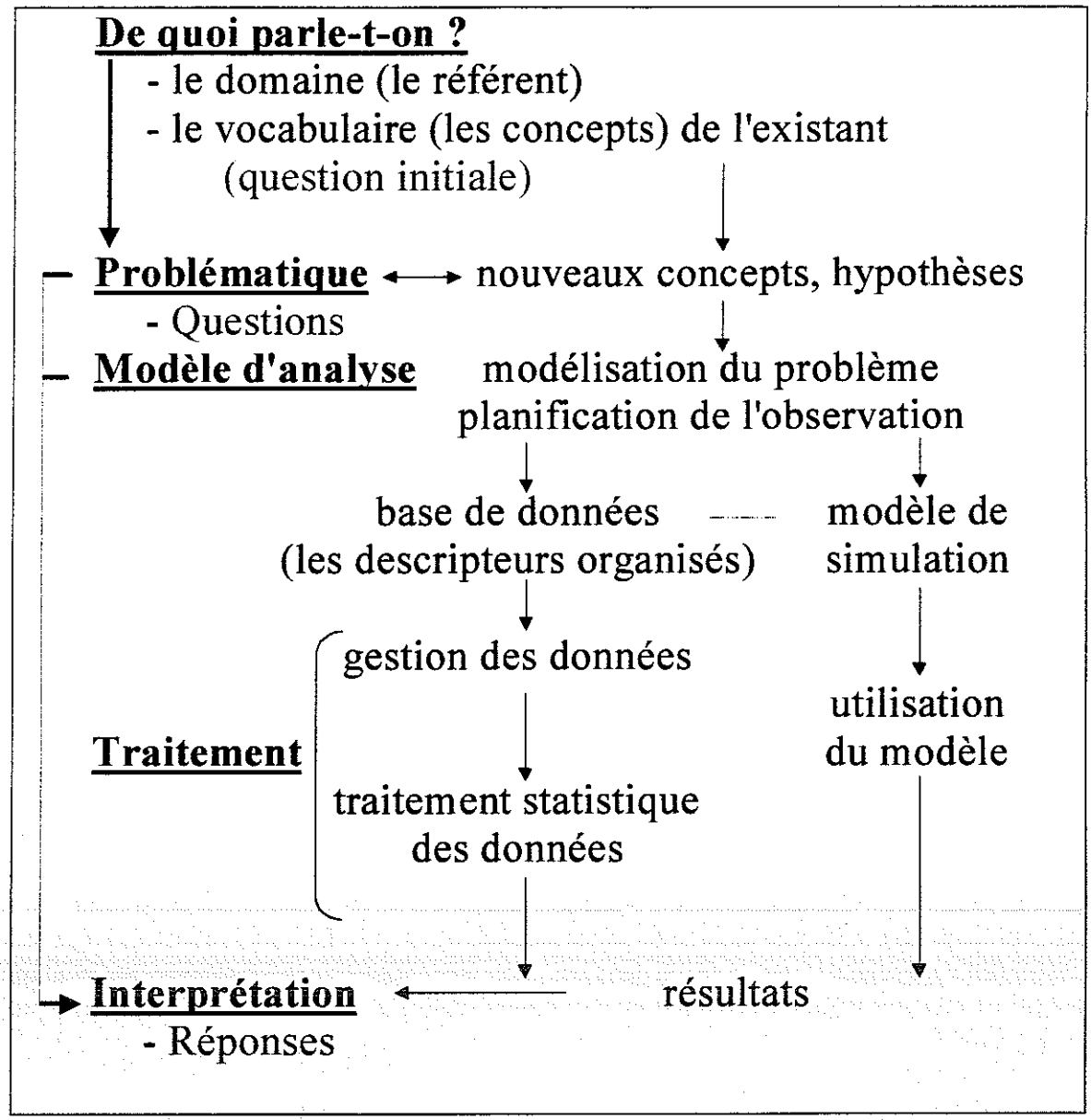


Figure 1 - La démarche méthodologique

A partir d'une question initiale et d'un travail préalable d'exploration de l'existant (étude d'une situation présente, de travaux antérieurs, de la bibliographie,...), la méthodologie d'approche des problèmes implique de passer par une étape au cours de laquelle une problématique pertinente et précise va apparaître [QUI95]. Dans le champ de l'étude, cette problématique s'exprime au moyen d'hypothèse(s) à vérifier, de concepts et d'hypothèses portant sur ces concepts. On rencontre généralement alors les questions sur le repérage de l'unité expérimentale, sur la hiérarchisation entre elles de notions génériques ou spécifiques, sur la prise en compte ou non de facteurs organisationnels par rapport à des principes invariants [COL87], indépendants de l'organisation choisie ensuite, ou encore sur le "maillage" à retenir dans l'espace ou dans le temps (échelles).
Ainsi, lors d'une expérience en élevage porcin, l'unité expérimentale sera à désigner précisément : s'agit-il de l'animal, de la loge, de la bande (groupes d'animaux élevés simultanément), ou de l'élevage ?

On aura alors construit et structuré un modèle qui, selon la nature du problème, est :
- un modèle d'analyse accompagné de procédés d'investigation (par exemple, une enquête) s'il s'agit d'observer des phénomènes ou de tester des hypothèses (ici, interviennent les méthodes et outils de la statistique),
- ou une modélisation de phénomènes (souvent au moyen d'équations exprimant des flux entre des sources et des puits) pour simuler un comportement ou des événements futurs (étude, par exemple, de la dynamique d'une population).

Et, en final, c'est bien cette problématique clairement formulée et concrétisée par un modèle qui permettra une interprétation réaliste des résultats obtenus, notamment en se référant aux hypothèses retenues ou formulées initialement.

2.1. La modélisation des concepts et des données

A l'instar de la démarche de conception d'une "application informatique" (ensemble des fichiers de données et des traitements associés à une fonction particulière), la phase amont d'études appuyées sur des données quantitatives consiste aussi à prendre en compte et définir les concepts retenus, et, en conséquence, à organiser entre elles les données : Ainsi, par exemple, l'approche du champ sémantique du problème abordé permettra – et imposera – de bien distinguer la notion de "groupes" de celle des "individus" qui composent ces groupes.
Un langage, particulièrement riche et concis, pour exprimer la réalité et les hypothèses retenues pourra être celui de la modélisation des données sous formes de types d'entités et de types d'associations entre ces types d'entités (appelée généralement "modélisation Entités/Associations" ([COL87] - pp. 32-60 , [GAL84] - pp. 31-44).

A titre d'exemple d'emploi de ce langage, on s'intéresse à une expérimentation conduite sur un verger composé de plusieurs parcelles. Pour simplifier l'exemple, on suppose que sur chaque parcelle les arbres sont tous de la même variété et qu'il n'est pas prévu de répétition des traitements. Le schéma (dit "schéma conceptuel") de la figure 2 est construit afin d'exprimer les caractéristiques retenues; il se lit ainsi :
- On considère une collection de parcelles expérimentales identifiées chacune par un numéro (exprimé par le type d'entité "Parcelle"). Sur chacune, on trouve plusieurs arbres, avec la possibilité qu'il n'y en ait cependant aucun (indiqué par les "cardinalités" minimum et maximum notées 0,n).
- Un arbre est localisé sur une parcelle : les cardinalités (1,1) précisent en effet que chaque arbre est localisé sur au moins une et sur au plus une parcelle.

- A diverses périodes, des traitements sont appliqués aux arbres ; les cardinalités 0,n placées du côté du type d'entité "Arbre" expriment qu'au cours du temps chaque arbre aura fait l'objet d'aucun traitement ou de plusieurs (à différentes périodes).
- A certaines périodes, pas forcément les mêmes, des mesures sont effectuées sur chaque arbre (par exemple : stade de floraison ou de fructification, nombre et calibre des fruits, présence de maladie,...).

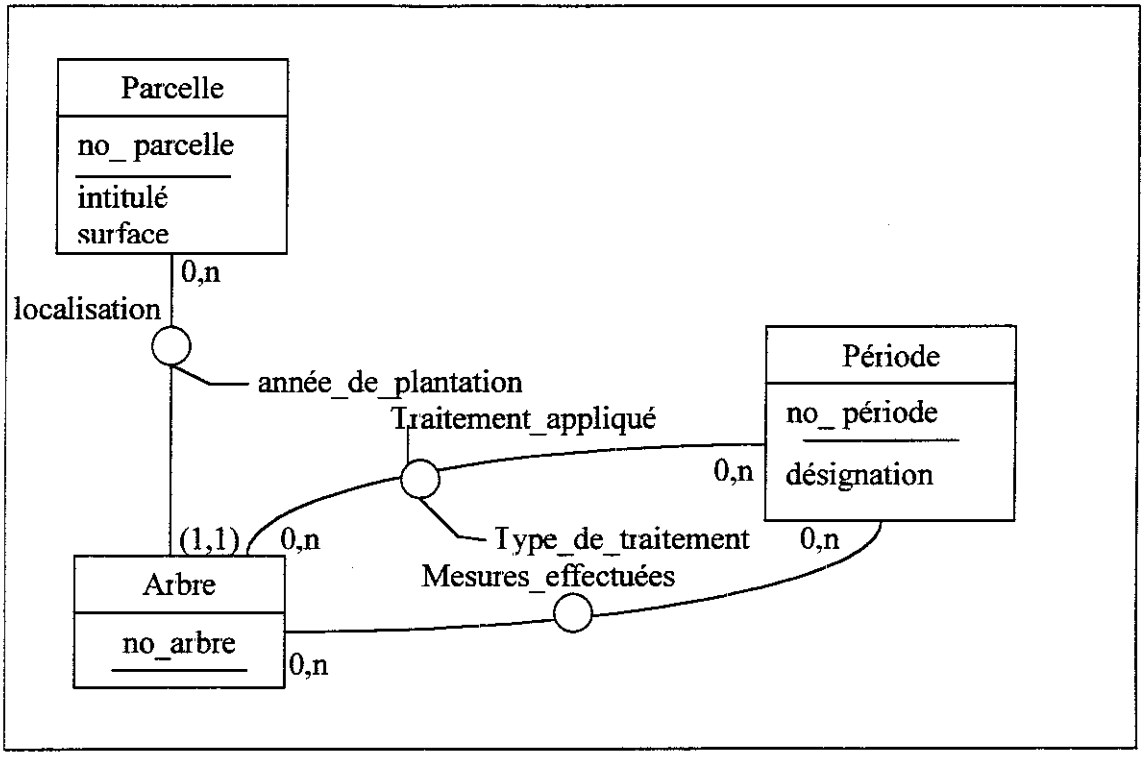


Figure 2 – Expérimentation sur verger : schéma conceptuel initial

Précisons le, ce formalisme ne sert pas seulement à exprimer la réalité retenue et à organiser ensuite les données. Son rôle premier est de contribuer de façon déterminante à l'émergence de la compréhension des concepts. Il est pleinement un langage pour l'analyse et pour une vision d'ensemble du domaine abordé :
Ainsi, pour le problème précédent, supposons que ce que l'on avait initialement compris de la réalité se trouve, un peu plus tard au cours de la démarche d'analyse, précisé par l'indication suivante :

- Lorsqu'on applique un traitement aux arbres, il est toujours appliqué à l'ensemble des arbres d'une parcelle.

Le schéma qui intègre cette information supplémentaire devient alors celui de la figure 3. Le lecteur habitué à la lecture du formalisme utilisé lira bien :

- Sur la figure 2 : un traitement différent peut être appliqué à chaque arbre, et à des périodes différentes.
- Sur la figure 3 : Un traitement concerne une parcelle (à une période donnée, il est identique pour chaque arbre de la parcelle).

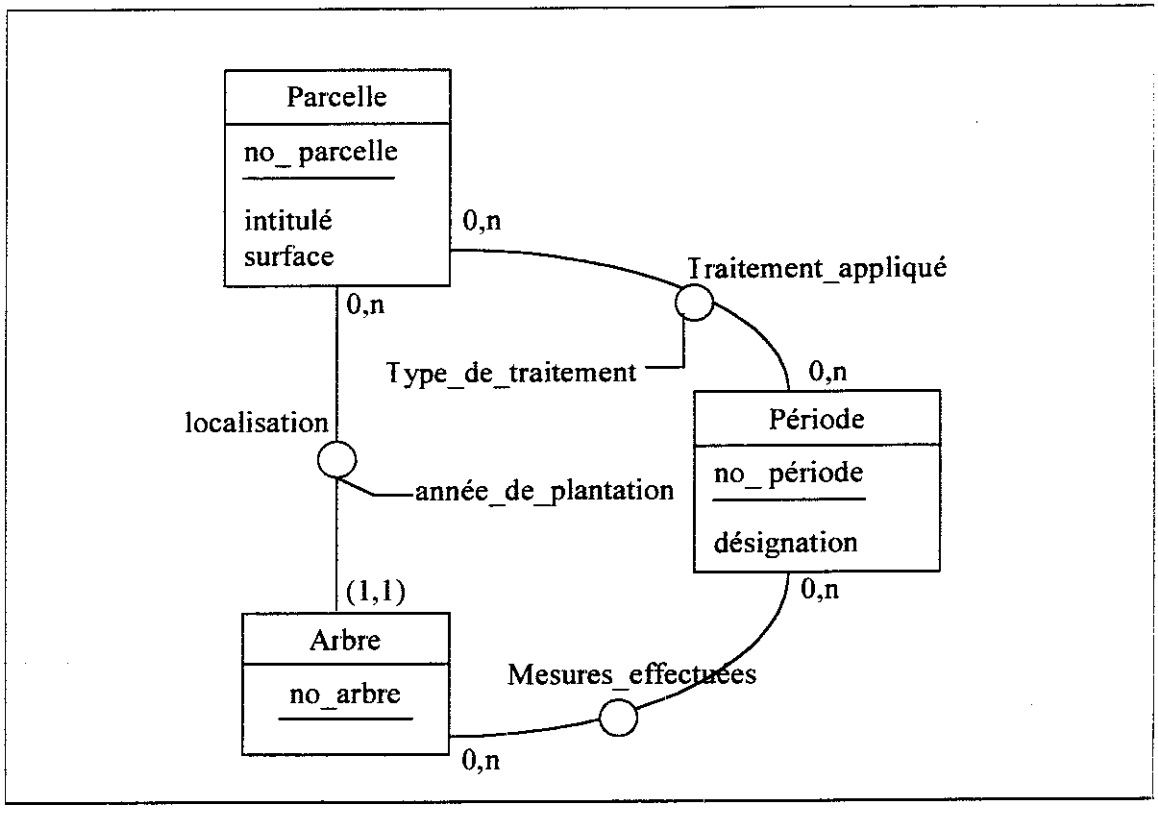


Figure 3 – Expérimentation sur verger : second schéma conceptuel

2.2. La structuration des données en une base de données

L'esquisse du schéma conceptuel de principe de la figure 3 devrait être complété ou précisé en indiquant comment décrire les types de traitements appliqués au moyen de propriétés relatives au type d'association "Traitement_appliqué" (par un simple numéro d'identification ou par la description des produits et doses employés, ou encore par les types de taille effectués) et quelles mesures sont précisément réalisées sur les arbres.

En découlera alors, par une simple opération de traduction, le "schéma physique" de la base de données, exprimé et normalisé selon les règles du modèle relationnel [CHR97] : ici, un ensemble de cinq tables (ou tableaux) reliées entre elles, contenant chacune les données relatives aux cinq populations de parcelles, d'arbres, de périodes, de traitements appliqués et de mesures effectuées. Et si l'on décide de gérer les données avec un SGBD (Système de Gestion de Bases de Données) - ce que suggère fortement le fait d'avoir des données organisées en base de données, mais ce qui n'exclue pas à priori l'usage d'un tableur -, l'allure de ce schéma physique, représenté sous sa forme graphique, est comparable à celui représenté sur la figure 4.

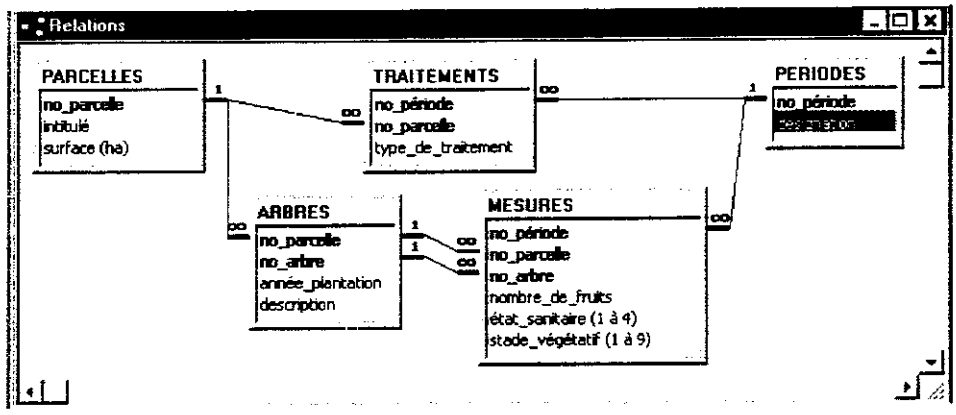
Remarquons tout de suite, mais nous y reviendrons : en suivant la description du problème tel qu'il se présente dans la réalité, les données relatives au verger expérimental sont organisées initialement dans cinq tableaux et non dans un seul !

Et constatons que, en ce qui concerne la structure de la base de données, et plus précisément celle de la table TRAITEMENTS, nous aboutissons à une structure différente de celle à laquelle aurait conduit le schéma conceptuel de la figure 2 :

à partir de la figure 3 : TRAITEMENTS (no_période, no_parcelle, type_de_traitement)

au lieu de, à partir de la figure 2 :

TRAITEMENTS (no_période, no_parcelle, no_arbre, type_de_traitement).



PARCELLES	no_parcelle	intitulé	surface (ha)
	1	parcelle 1	1
	2	parcelle 2	1.5

ARBRES	no_parcelle	no_arbre	année_pl	description
	1	1	1988	arbre 1 de la parcelle 1
	1	2	1988	arbre 2 de la parcelle 1
	1	3	1989	arbre 3 de la parcelle 1
	2	1	1990	arbre 1 de la parcelle 2
	2	2	1990	arbre 2 de la parcelle 2

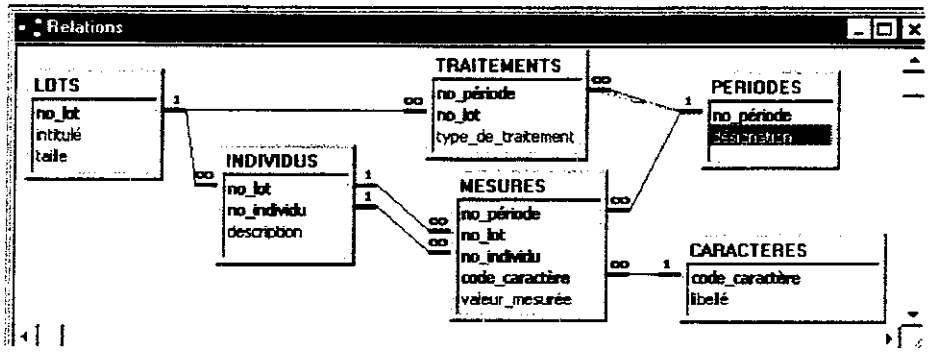
PERIODES	no_période	désignation
	1	début avril
	2	début mai
	3	début juin

TRAITEMENTS	no_période	no_parcelle	type_de_traitement
	1	1	3
	1	2	2
	2	1	5

MESURES ---------- ci-dessous, les caractères mesurés :

no_période	no_parcelle	no_arbre	nb_fruits	état_santé	stade_végét
1	1	1	0	1	2
1	1	2	0	2	3
1	1	3	0	2	2
.....	.....	.....	.....	.....	.....
2	1	1	0	1	6
2	1	2	0	3	7

Figure 4 - Expérimentation sur un verger : schéma physique



LOTS	no_lot	intitulé	taille
	1	parcelle 1	1
	2	parcelle 2	1.5

INDIVIDUS	no_lot	no_individu	description
	1	1	arbre 1 de la parcelle 1
	1	2	arbre 2 de la parcelle 1
	.....	.....	.....

Tables PERIODES et TRAITEMENTS comparables à celles de la figure 4

CARACTERES	code_caractère	libellé
	NBF	nombre de fruits
	SNT	état sanitaire (noté de 1 à 4)
	STV	stade végétatif (noté de 1 à 9)

MESURES ---------- caractères mesurés :

no_période	no_lot	no_individu	code_caractère	valeur_mesurée
1	1	1	NBF	0
1	1	1	SNT	1
1	1	1	STV	2
1	1	2	NBF	0
1	1	2	SNT	2
1	1	2	STV	3
.....	.....	.....	.....	.....
2	1	1	NBF	0
2	1	1	SNT	1
2	1	1	STV	6

Figure 5 - Expérimentation sur lots d'individus : schéma physique

2.3. La généralisation du problème et le paramétrage

Selon l'exemple présenté antérieurement, la figure 4 montre donc la structuration des données et le contenu des tables que l'on peut envisager pour la réalisation d'expérimentations effectuées sur les parcelles d'un verger.

Et la figure 5 montre l'idée d'une généralisation possible du logiciel : En l'étendant à des expérimentations portant sur tout ensemble de lots d'individus sur lesquels seront effectuées, à des périodes déterminées, des mesures. Ces mesures, évidemment, dépendent alors du type d'individus concernés et seront, dans cette seconde version de logiciel, à décrire en tant que données (voir tables CARACTERES et MESURES de la figure 5) et non plus en tant que caractères prédéfinis (comme dans la table MESURES de la figure 4).

Et, avec ce paramétrage préalable, voici qu'au lieu du logiciel spécifiquement bâti pour traiter les données d'expériences sur un verger, apparaît une esquisse de ce que peut être un logiciel général d'expérimentation.

3. Justification et avantages de l'approche préconisée

Dans ce chapitre, nous présentons à la fois la justification des principes et l'intérêt des possibilités que nous venons de passer en revue :

3.1. La structuration de la base de données en deux étapes

Si on considère de nouveau le schéma général de la démarche d'une étude proposé sur la figure 1, on perçoit bien que l'expression de la problématique et la modélisation associée engendrent ensuite une organisation des données. Cette organisation de descripteurs constitue d'une part la formulation sémantique du problème et permet d'autre part, lors des traitements, le maniement syntaxique des valeurs affectées à ces descripteurs.

Aussi, et surtout pour les problèmes les plus complexes, il est déterminant de bien distinguer les deux étapes qui doivent se succéder :

- celle d'abord, comme on l'a montré en détail sur l'exemple du verger, qui permet à la fois de repérer et de structurer entre eux les concepts du problème ("schéma conceptuel" décrit au moyen de la modélisation en Entités/Associations),
- puis celle de traduction en "schéma physique" de base de données.

Car, si on a respecté les règles, peu nombreuses, régissant d'une part la construction du schéma conceptuel et d'autre part l'opération de traduction en tables et relations, on obtient alors une structure de base de données normalisée selon les règles du modèle relationnel. Cette normalisation garantit deux choses :

- la non-redondance des données entre elles (caractéristique à rechercher impérativement),
- et par voie de conséquence, l'aptitude des données à être manipulées et traitées correctement par le langage normalisé (SQL : Structured Query Language) résidant au cœur des logiciels de type tableur ou SGBD, c'est-à-dire, pratiquement, une facilité pour la formulation de variables calculées, l'extraction ou la construction de tableaux, l'obtention de résultats statistiques...

En réalité, vouloir analyser un problème complexe tout en construisant directement la structure physique des données est une démarche forcément vouée à l'échec !

Au passage, nous venons d'évoquer tableur et SGBD... Est-il besoin d'attirer l'attention sur le fait que l'appellation "base de données" est employée ici selon sa définition ("ensemble de tables reliées entre elles") et non dans le sens que lui donne le tableur EXCEL (une seule table !) ?

3.2. Un niveau de généralité adéquat

Dépassant les seuls aspects pratiques du type de logiciel mis en œuvre et du développement informatique, la notion de "système d'information" englobe le continuum formé de l'organisation autour de ce logiciel (modalités de collecte, de saisie, de traitement, de diffusion des résultats) et des procédés, parfois complexes, régissant cette organisation et les fonctionnalités de ce logiciel [GAL84] [MEL85].

Nous évoquons ici la nécessité de disposer d'un système adapté aux finalités d'une organisation du traitement de l'information, dont la composante humaine n'est pas exclue, nous allons le montrer.

L'une des manières de réussir cette adaptation des logiciels aux finalités qu'on leur assigne est celle retenue par les concepteurs de "progiciels". Selon le Journal Officiel du 7 décembre 1980, le progiciel est en effet un "produit-logiciel" : conçu comme un produit industriel destiné à une diffusion en masse et accompagné de services [CXP]. Ainsi, on le sait, un progiciel de comptabilité n'est pas développé pour une seule entreprise mais pour les entreprises, un progiciel de gestion technique d'un élevage est destiné à tout éleveur.

Comment adapter alors le progiciel à telle entreprise ou à tel éleveur de telle sorte qu'ils apparaissent à leurs utilisateurs comme exactement dimensionnés à leurs besoins ?

C'est par l'intermédiaire d'un procédé de "paramétrage" que l'utilisateur va pouvoir communiquer ses spécificités (son plan comptable par exemple). Et lorsque plus tard des conditions ou des "règles de gestion" varieront, la modification possible des valeurs des paramètres garantit une meilleure pérennité du progiciel par rapport à celle d'un logiciel conçu pour ne fonctionner que dans le contexte d'un environnement particulier et à un moment donné.

C'est bien ce qui a été fait avec l'esquisse de logiciel d'expérimentation introduite au chapitre précédent : un seul outil apte à traiter des problèmes différents et, pour un même problème, capable d'intégrer une évolution possible au cours du temps des caractères à mesurer.

Toutefois, pour un logiciel apte à prendre en compte toutes les composantes de la réalité - et non plus simplifié pour des raisons didactiques -, les concepts abstraits autorisant un fort degré de paramétrage peuvent devenir suffisamment complexes pour empêcher son emploi par un utilisateur non spécialement formé et averti des fonctionnalités potentielles de l'outil, ... ou ayant du mal à les comprendre. L'équilibre donné par un niveau de généralité adapté est donc à raisonner aussi en fonction de cet utilisateur !

3.3. Les données ne sont pas a priori rassemblées en un seul tableau !

Et maintenant, quels arguments fournir au statisticien qui organiserait systématiquement les données d'une étude en un seul tableau, quels arguments pour l'inciter à ranger les données initiales dans autant de tableaux que de types d'entités repérés ? Nous les fournirons en trois points :

- D'abord en rappelant que les logiciels, SGBD ou tableur, disposent de commandes qui permettent d'obtenir, à partir des données brutes organisées comme on l'a suggéré, toutes formes de tableaux souhaités pour le traitement (que ce traitement soit réalisé avec un tableur ou avec un logiciel statistique spécifique).

Ainsi, si on a réalisé des pesées sur des animaux à diverses périodes (donc, trois tableaux : les animaux, les dates de pesée et les mesures effectuées), les poids observés doivent être enregistrés dans une table de mesures telle que celle apparaissant sur la figure 6.

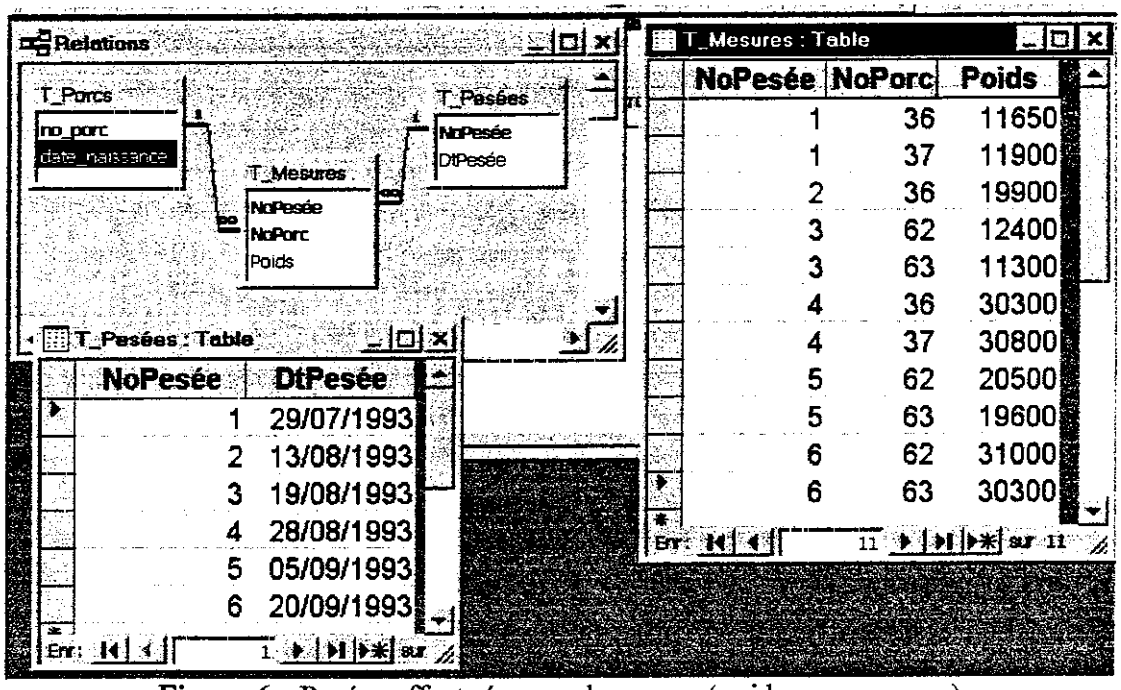
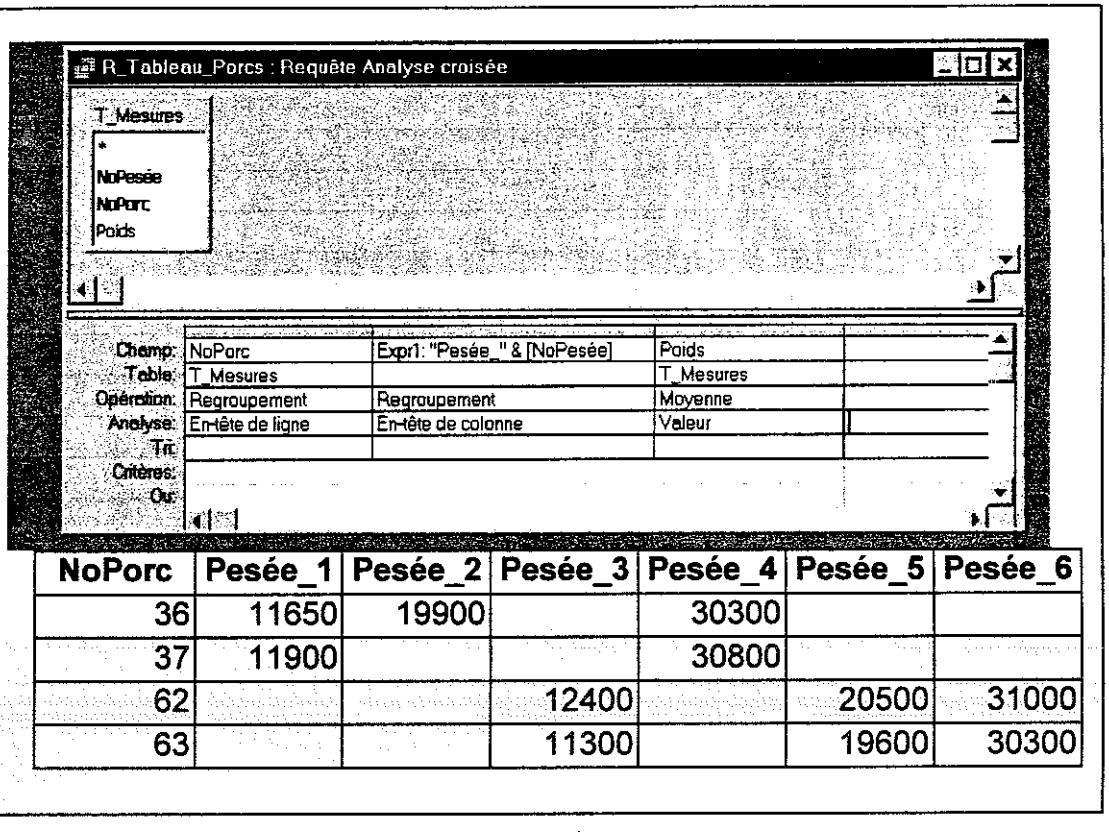


Figure 6 - Pesées effectuées sur des porcs (poids en grammes)



NoPorc	Pesée_1	Pesée_2	Pesée_3	Pesée_4	Pesée_5	Pesée_6
36	11650	19900		30300		
37	11900			30800		
62			12400		20500	31000
63			11300		19600	30300

Figure 7 - La requête permettant de retrouver un tableau à double entrée

La figure 7 montre la rédaction sous ACCESS de la requête simple qui s'applique à cette table et permet d'obtenir le tableau à double entrée qu'on aurait pu souhaiter créer d'emblée..... ce tableau peut ensuite être exporté vers le tableur.

- Mais supposons que sur un tel tableau, au sein duquel on remarque la présence de valeurs manquantes, on souhaite obtenir la succession des accroissements de poids entre toutes les périodes consécutives où les valeurs sont présentes : voir sur la figure 8 les résultats attendus (l'accroissement est exprimé ici par le GMQ : Gain de poids Moyen Quotidien). Pour l'individu n°36, ces GMQ sont calculés entre les pesées 1 et 2 d'une part, 2 et 4 d'autre part puisque le poids est manquant à la pesée 3. Pour un nombre important de données, l'automatisation de ce calcul avec une feuille de tableur est difficilement réalisable : elle nécessite une écriture en langage procédural accompagnant la feuille de calcul. Tandis que, si les données se trouvent sous la forme originelle de la figure 6 (forme déduite du schéma conceptuel), une requête seule répond à la question : voir, figure 8, cette requête, certes moins facile à composer que la précédente.

NoPorc	np0	d0	p0	np1	d1	p1	GMQ
36	1	29/07/199	11650	2	13/08/1993	19900	550.00
36	2	13/08/199	19900	4	28/08/1993	30300	693.33
37	1	29/07/199	11900	4	28/08/1993	30800	630.00
62	3	19/08/199	12400	5	05/09/1993	20500	476.47
62	5	05/09/199	20500	6	20/09/1993	31000	700.00
63	3	19/08/199	11300	5	05/09/1993	19600	488.24
63	5	05/09/199	19600	6	20/09/1993	30300	713.33

avec : np_i = n° de pesée / d_i = date de la pesée p_i = poids mesuré /
et i = 0 ou 1 (début ou fin)
GMQ : gain de poids moyen quotidien = $(p_1-p_0)/(d_1-d_0)$

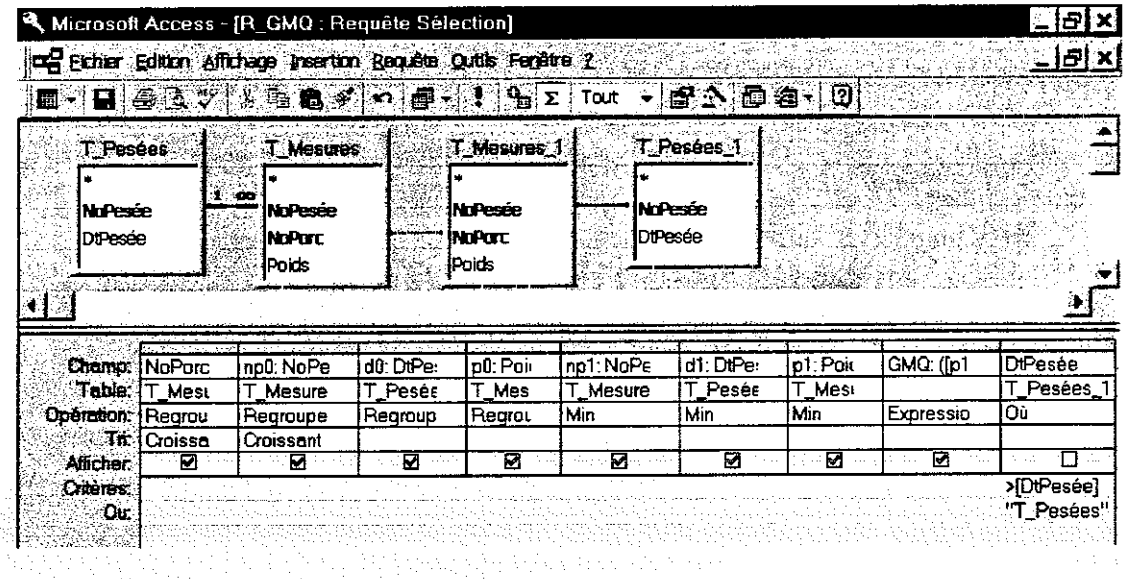


Figure 8

Requête de type réflexif automatisant le calcul des accroissements entre dates successives

- Enfin, bien sûr, détenir séparément dans plusieurs tables les données des diverses populations concernées évitera de calculer par inadvertance une moyenne pondérée sur des valeurs rapportées plusieurs fois dans un tableau (par exemple, lors du calcul de la surface moyenne des parcelles à partir d'un tableau contenant, pour chaque arbre, les caractères le concernant et ceux de la parcelle à laquelle il appartient).

Donc, si le statisticien pense parfois d'abord à la structure du tableau qu'il veut traiter, ... une structure comparable par exemple à celle-ci :

		< ----- Période 1 ----- >			< ----- Période 2 ----- >			< -----
no_parcelle	**no_arbre**	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{31}

il doit cependant s'interdire de penser que le stockage des données de base doit être effectué sous cette forme !
Car alors, comment ferait-il s'il voulait faire calculer systématiquement des accroissements de valeurs entre les périodes où elles sont effectivement renseignées ? Et aussi, comment ferait-il si, stockant ces données au cours de plusieurs années (en ayant rajouté en première colonne le millésime), la réglementation concernant la qualité des aliments imposait à partir d'un certain moment de mesurer des caractères supplémentaires ?
Seule, la structure physique des données découlant de la représentation de la réalité (en plusieurs tables) le lui permettra, le protégeant des incohérences et redondances entre les données tout en laissant possible à tout moment la construction du tableau souhaité pour le traitement.

3.4. Intérêt du recours à un SGBD

L'aboutissement de la démarche d'analyse et d'organisation des données débouche naturellement sur une structure en base de données ...
Nous énumérons alors maintenant les raisons qui nous font préférer un SGBD à un tableur pour le stockage et la gestion des données collectées (le tableur pouvant ensuite être utilisé pour la réalisation de calculs et de graphiques). Ces raisons sont essentiellement celles relatives à la sécurité et à la commodité d'emploi :

- Intégrité des individus assurée ! Un SGBD gère en effet un "enregistrement" relatif aux données d'un seul individu : une fois quitté, après toute modification, cet enregistrement est immédiatement (et automatiquement) écrit sur la mémoire de masse. Et la demande d'un tri pour visualiser les individus dans un certain ordre ne risque pas de conduire à une catastrophe comme celle qui se produit avec un bloc de cellules de tableur mal sélectionnées : le SGBD, là aussi, trie d'office les enregistrements, et non certaines rubriques de ces enregistrements !
- Etablissement des liaisons entre les tables très aisé : avec la souris, on relie les champs (caractères) qui se correspondent (voir relations sur les figures 4 et 5) ; sont associées à chaque liaison les valeurs des cardinalités qui permettent de demander au logiciel d'effectuer lui-même les contrôles d'intégrité lors de la saisie (pour un arbre, situation effective sur une seule parcelle et sur une parcelle existante !).
- Très grande facilité pour créer des masques de saisie permettant, par exemple, de saisir "simultanément" une période et les mesures réalisées à ce moment là sur les arbres (voir figure 9).

- Et enfin, mise à disposition d'aides visuelles pour la rédaction des requêtes ("requêteurs", voir figure 8) et, lors de leur exécution, exploitation systématique des divers types de jointures décrites entre les tables (jointures strictes ou jointures externes pour prendre en compte l'existence d'enregistrements manquants).

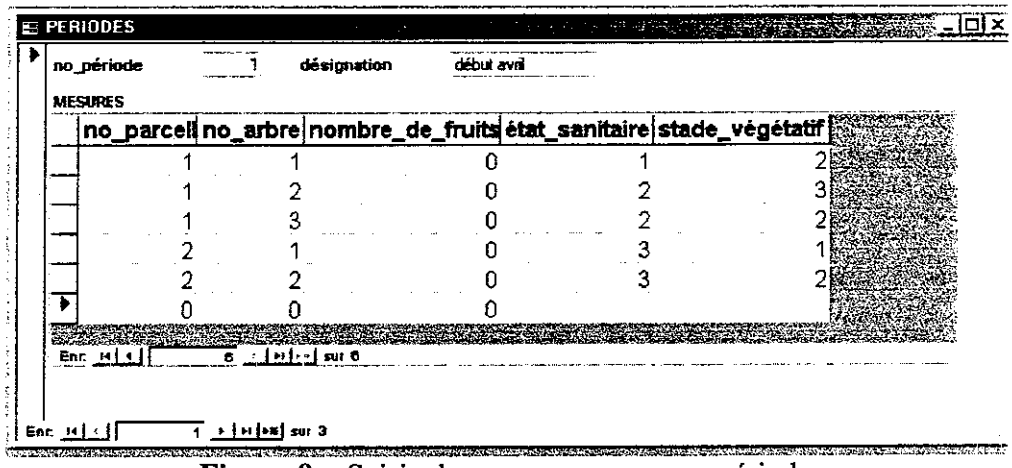


Figure 9 – Saisie des mesures pour une période

Ces avantages du SGBD nous font dire que bientôt le tableur n'apparaîtra plus, comme il y a dix ans, l'outil presque incontournable pour la gestion des données en utilisation individuelle. Mais, disant cela, il nous faut bien confirmer que l'indispensable approche du problème sous forme du schéma conceptuel n'interdit pas cependant, à celui qui le préférerait, d'organiser les données dans des feuilles de tableur,... pour peu qu'il n'y ait pas un trop grand nombre de relations à gérer et que soient développées des fonctionnalités assurant l'intégrité des données entre elles!

4. Modèles de données pour les questionnaires

En appliquant les principes qui viennent d'être abordés, nous discuterons deux cas :

- celui de l'enquête occasionnelle, ponctuelle, débouchant sur une collection de formulaires remplis, parfois de façon anonyme,
- et celui du panel : informations collectées à diverses périodes auprès d'individus identifiés, avec la possibilité de choisir les questions posées d'une fois sur l'autre, et le souhait de rapprocher entre eux les résultats aux différentes périodes.

4.1. L'enquête occasionnelle

Un progiciel de dépouillement d'enquêtes s'appuie forcément sur une structure de base de données pourvue d'une possibilité de paramétrage permettant à l'utilisateur de fournir successivement deux types d'informations : dans un premier temps la description des questions posées, puis les réponses collectées sur les formulaires d'enquête.

Ainsi, se présente l'alternative suivante :

- Tel ou tel progiciel de dépouillement d'enquête conviendra-t-il pour ce que nous voulons faire ?
- Ou bien, allons nous construire une base de données dont la structure est strictement adaptée aux questions posées ? C'est ce qui a été fait pour l'enquête sur les logiciels de Statistique conduite en 1996 par l'ASU [ASU96] ; la figure 10, extraite du compte rendu de cette enquête ([ASU96] - p.8), redonne la structure de cette base, avec repérage des codes de quelques-unes des questions. On remarque que les concepteurs ont dû parfois

créer une table dédiée uniquement aux réponses à une seule question à choix multiples (exemple : question Q18).

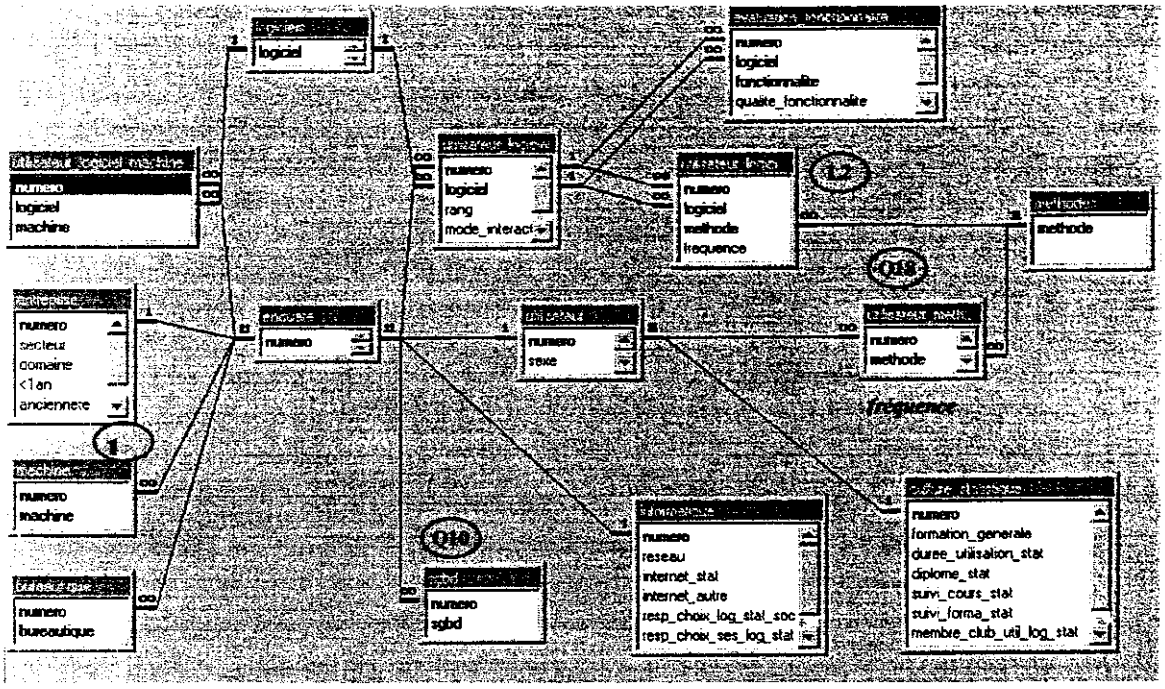


Figure 10 – Base de données pour l'enquête ASU

Mais, une troisième possibilité se présente : celle de l'option intermédiaire consistant, si le problème s'y prête, à construire soi même une base de données de type plus général. Le schéma de principe en est donné sur la figure 11, avec, toujours pour l'enquête qui vient d'être évoquée, des exemples pour le paramétrage des questions dans les tables QUESTIONS et TEXTES_ITEMS représentés sur la figure 12 (attention : apparaissent sur cette figure les informations, saisies par l'utilisateur du logiciel, selon une représentation de principe montrant le lien existant entre les questions et les items qui leur sont associés). Des exemples de réponses saisies dans la table REPONSES sont montrées sur la figure 13 (ces réponses ont pu être saisies peut-être par l'usager enquêté même, par exemple au moyen de formulaires HTML sur un site du "web").

Alors, avec quelques requêtes générales, on peut obtenir les résultats des réponses à l'ensemble des questions : ainsi, figure 14, toutes les valeurs moyennes des réponses aux questions de type quantitatif (question Q0) et, figure 15, toutes les fréquences des modalités de réponses aux questions de type qualitatif données par les textes d'items (question Q10). Une requête proche de la précédente fournira les mêmes résultats pour les réponses aux questions de même type données par des valeurs (question Q18).

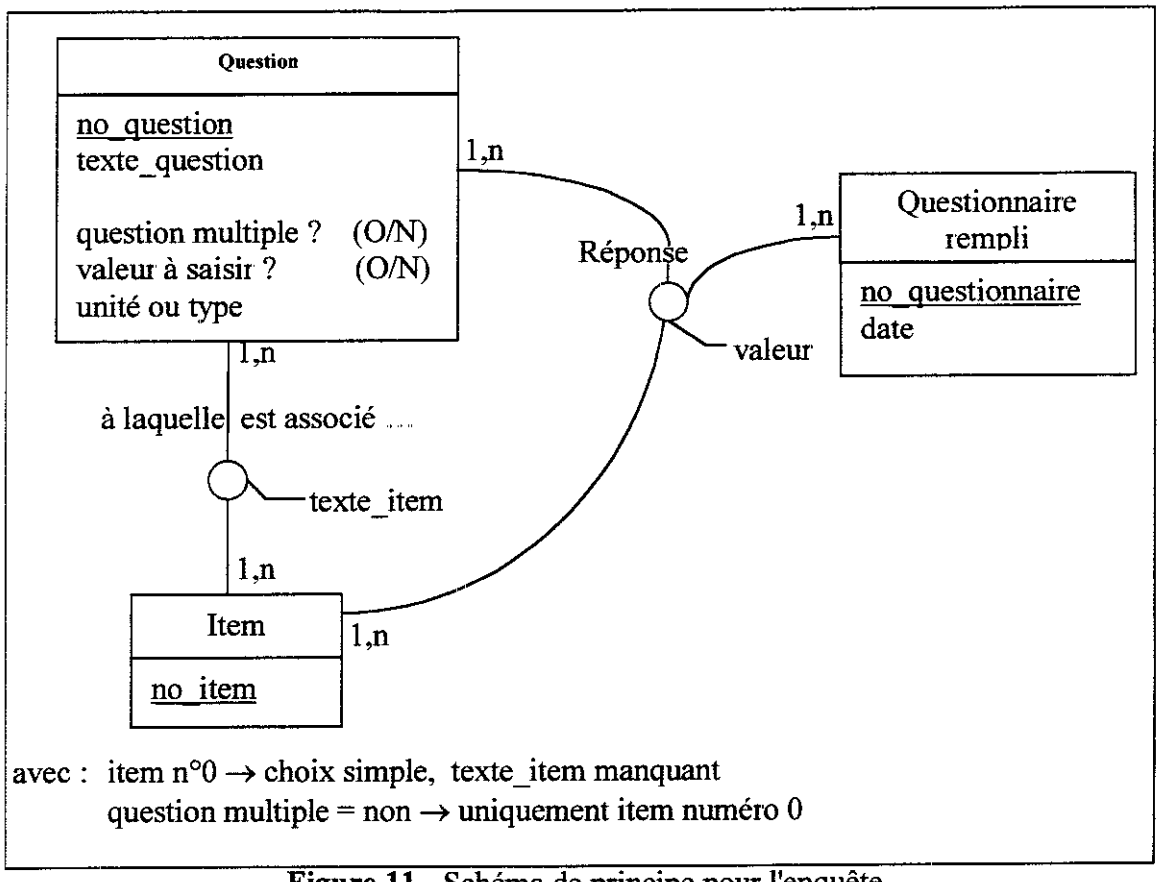


Figure 11 - Schéma de principe pour l'enquête

QUESTIONS				**ITEMS et TEXTES ITEMS**				
N°	**Texte**	**Question multiple ?**	**Valeur à saisir**	**0**	**1**	**2**	**3**	**4**
Q0	Age	N	O	/				
Q6	machine	O	N		PC	Mac	Station	Mini
Q10	SGBD	O	N		Paradox	Oracle	Access	
Q18	Fréquence méthode	O	O		Comptage	Statistiques descriptives	Régression linéaire	Analyse de la variance
Q19	Logiciel de rang	O	O		Rang 1	Rang 2	Rang 3	Rang 4
L2-1	Fréquence méthode avec logiciel de rang 1	O	O		Comptage	Statistiques descriptives	Régression linéaire	Analyse de la variance
L2-2	Fréquence de la méthode avec logiciel de rang 2	O	O		Comptage	Statistiques descriptives	Régression linéaire	Analyse de la variance

Figure 12 - le paramétrage des questions

numéro de questionnaire	numéro de question	numéro d'item	valeur
123	Q0	0	50
123	Q6	1	
123	Q6	3	
123	Q10	2	
123	Q10	3	
123	Q18	1	3
...	...	...	...
123	Q19	1	SAS
123	Q19	2	StatGraphics
...	...	...	...
123	L2-1	1	2
123	L2-1	2	3
.....	.....	.....	.....
124	Q0	0	40

Figure 13 - Exemples de réponses possibles

Tables requises : REPONSES, QUESTIONS, ITEMS, TEXTES_ITEMS

Champ	no_q	texte_q	à_saisir?	unité	no_item	Texte_it	valeur
Table	REPONS	QUEST	QUEST	QUEST	REPONS	TEXTES	REPONS
Opération	Regroup.	Regroup.	Où	Où	Regroup.	Regroup.	Moyenne
Critère			O	<> Null			

Figure 14 - Requête pour obtention des moyennes (sous ACCESS)

Tables requises : REPONSES, QUESTIONS, ITEMS, TEXTES_ITEMS

Champ	no_q	texte_q	à_saisir?	unité	no_item	Texte_it	no_questionnaire
Table	REPONS	QUEST	QUEST	QUEST	REPONS	TEXTES	REPONS
Opération	Regroup.	Regroup.	Où	Où	Regroup.	Regroup.	Compte
Critère			N	Null			

Figure 15 - Requête pour obtention des fréquences de modalités (sous ACCESS)

De plus, une telle structure permet de réutiliser à la fois la base de données et les traitements (écrans de saisie, requêtes et états fournissant les résultats) si l'on décide, en cas de reconduction de l'enquête, de supprimer ou de rajouter des questions.

Reste à vérifier, si on s'oriente dans une telle direction, si un progiciel n'aurait pas fait l'affaire, ceci de façon plus économique en pensant au temps que l'on va consacrer à cette réalisation. En tous cas, cette troisième option aura montré l'intérêt d'une approche qui ne soit pas trop dépendante de la vision que l'on a, à un moment donné, d'un problème, alors que cette vision est appelée à évoluer, peut-être déjà en constatant les premières réponses obtenues lors du dépouillement.

4.2.Le panel

La question d'un logiciel général pour les panels s'est posée à l'ISAB pour gérer les nombreux documents de collecte d'informations observés dans les organismes et groupements agricoles : voir des exemples sur la figure 16.

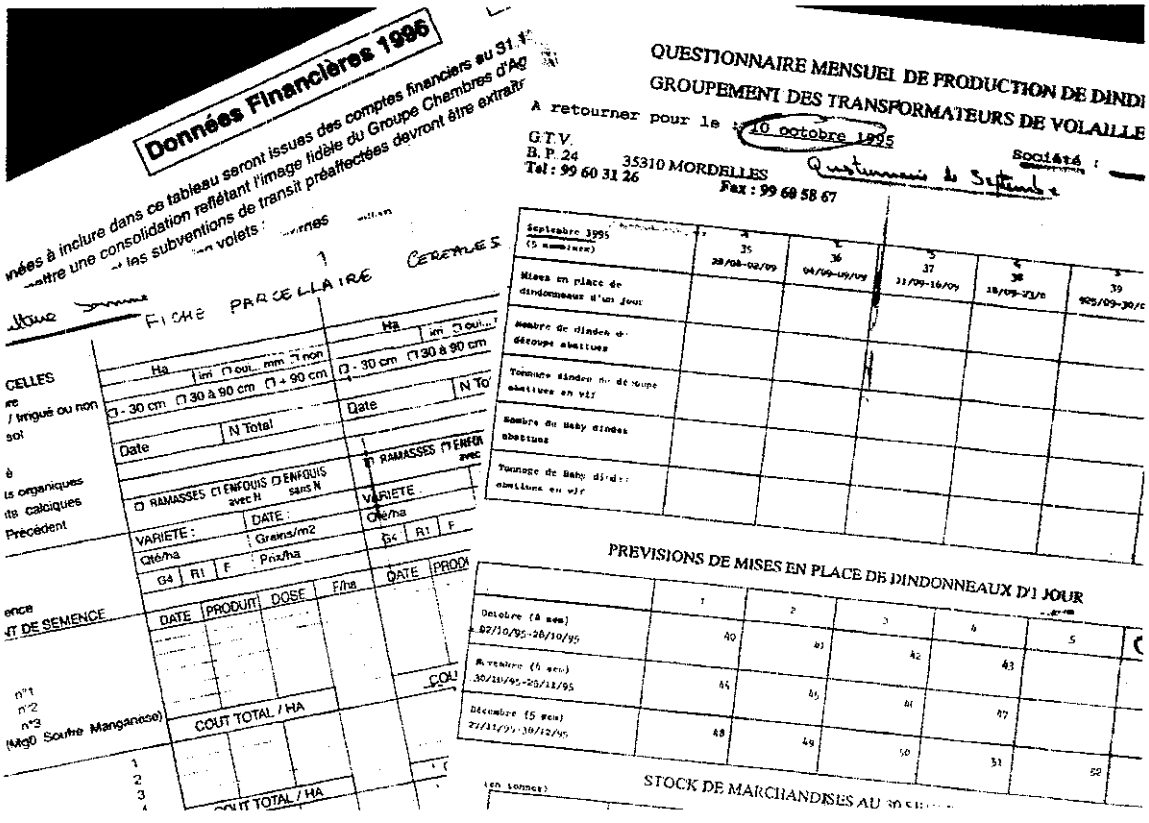


Figure 16 - Exemples de documents de collecte

Nous avons d'abord envisagé une base de données sur le principe évoqué dans le paragraphe précédent : les questions posées sont constituées chacune du croisement d'une "rubrique" avec un "item", deux tables de la base de données étant prévues pour la description des ensembles de rubriques et d'items. Cette organisation permet de prendre en compte les questions présentées dans des tableaux à double entrée comme sur le formulaire apparaissant au premier plan de la figure 16 (activité mensuelle de producteurs de dindes). La figure 17 montre, pour ce même formulaire, le paramétrage des rubriques, des items, et donc des questions (au croisement des deux).

Actuellement, et pour bien prendre en compte les questionnaires les plus complexes, la réflexion nous conduit à préconiser une structure de principe telle que celle représentée sur le schéma conceptuel de la figure 18. Cette organisation permet en effet d'intégrer le fait que le panel est formé d'entités composées (par exemple, entreprises ayant divers sites géographiques), et de prendre en compte des formulaires composés de fiches reproduites n fois (des exploitations agricoles et leurs parcelles, par exemple).

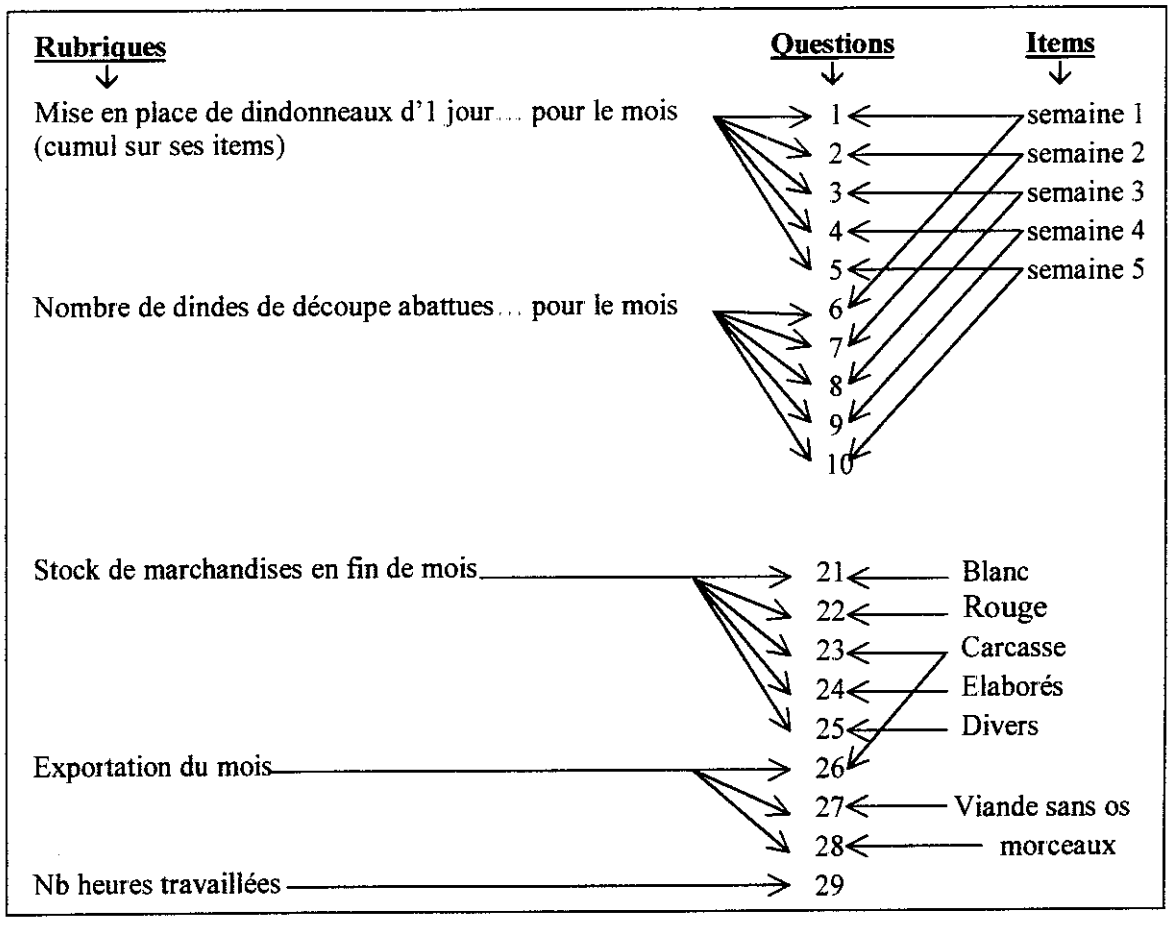


Figure 17 - Les questions apparaissant dans des formulaires à double entrée (rubriques X items)

Il est à noter que plusieurs autres concepts, non dessinés sur ce schéma de principe, sont encore nécessaires pour rendre complètement exploitable ce système. Du coup, son développement, forcément dans la perspective d'un progiciel, suppose un niveau d'investissement important (après avoir bien vérifié qu'il n'y a pas de produit existant qui couvrirait l'ensemble des fonctionnalités souhaitées). Cet effort pourrait alors être réparti entre plusieurs partenaires s'unissant (l'équipe de la cellule STID de l'ISAB pouvant faire partie du noyau de conception, en amont de la réalisation proprement dite). Nous considérons en effet que le développement d'un tel outil devient maintenant décisif quand on pense aux demandes décuplées en matière de recueil d'informations auprès de partenaires ou de prospects (développement du commerce électronique via les "sites web" interactifs dont beaucoup d'entreprises ou d'organismes souhaitent se doter).

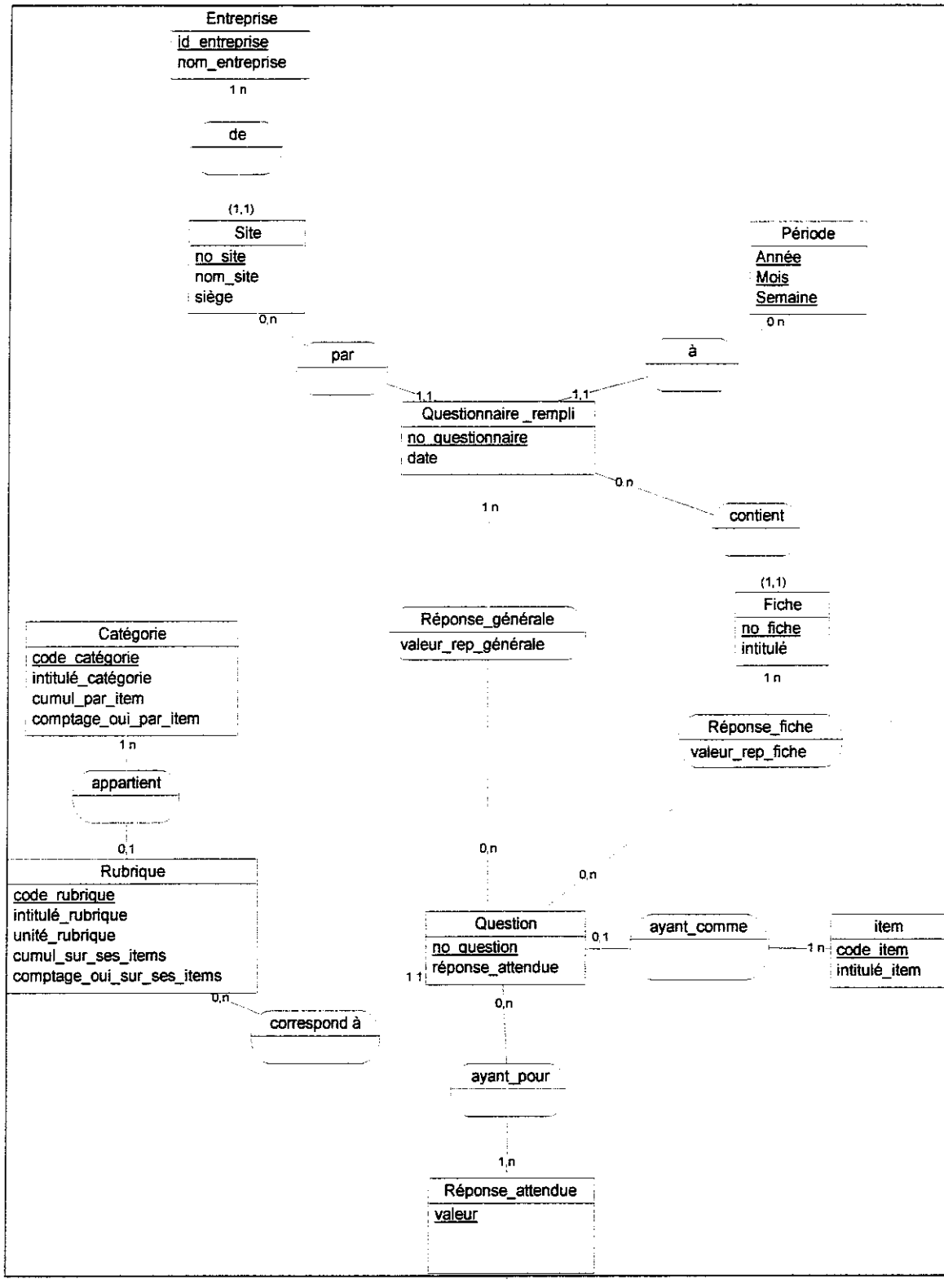


Figure 18 - Panel : schéma conceptuel de principe

5. Le problème des questions hiérarchisées

Que ce soit pour conserver des mesures résultant d'expériences ou des réponses obtenues au moyen de questionnaires, le recours à une organisation de rubriques hiérarchisées entre elles est souvent souhaité. Ce sera le cas pour des questions à réponses multiples, souvent elles-mêmes conditionnées par la réponse à une question générique, comme par exemple :

- présence d'animaux dans le logement (oui/non) ?
- si oui, nombres de chiens ? ... de chats ? ... d'autres ?

De telles informations hiérarchisées sont aussi appelées "variables ou questions mère-fille" ([STE99], p12).

D'une part, l'organisation des rubriques (caractères mesurés ou questions posées) ne peut alors être conçue que de la manière dont on l'a préconisé au paragraphe 2.3 et dans le chapitre précédent sur la représentation des questionnaires : en les décrivant dans une table en tant que données (et non comme les champs d'une table).
Et d'autre part, la hiérarchie existant entre elles est à envisager comme celle rencontrée dans les systèmes utilisant les techniques de classification :

- Thésaurus (exemple : TG → genre humain TS → homme, femme
  avec TG = Terme Générique et TS = Terme Spécifique).
- Taxonomies (exemples : familles de plantes, regroupant des genres, constitués eux-mêmes de diverses espèces, ou catégories socio-professionnelles décomposées en sous-catégories).

A partir de là, deux démarches sont envisageables pour représenter ce qui est en fait une hiérarchie sous forme d'arborescence :

5.1.La "structure réflexive"

Il s'agit de celle utilisée pour décrire des nomenclatures (tel composant, constitué lui-même de plusieurs composants, entre dans la composition de tel composant); cette structure est suggérée par les schémas de la figure 19 : items d'une taxonomie, composants d'une nomenclature, questions d'un questionnaire.

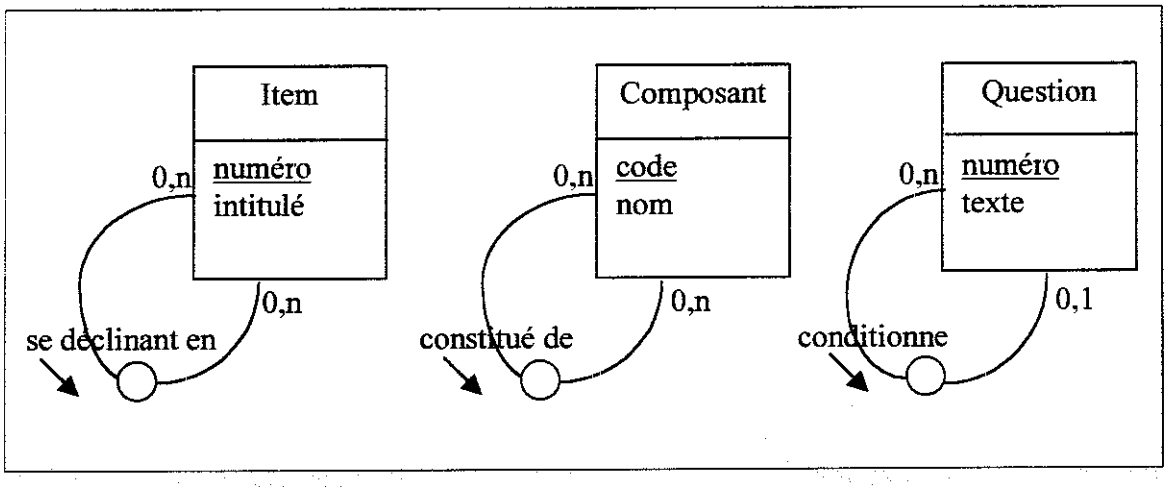


Figure 19 : Structures réflexives

La description de questions conduit à un rangement dans la table QUESTIONS semblable à celui-ci :

Numéro_question	Texte de la question	Conditionnée par question n°
...		
47	Présence d'animaux (Oui/Non) ?	
48	Nombre de chiens ?	47
49	Nombre de chats ?	47

Une structure de ce type devra être accompagnée d'une procédure de type récursif pour le parcours de l'arborescence des questions ainsi emboîtées sur autant de niveaux qu'on le souhaite :

Procédure Recherche_questions-filles (Nq) (avec : Nq → Numéro_question)
Requête donnant les numéros des questions dont Cp = Nq
(avec : Cp → valeur de la rubrique "Conditionnée_par_question_n°")
UNION :
Recherche_questions-filles (pour les numéros que l'on vient d'obtenir)
Fin de procédure

5.2.L'emploi d'un code décimal hiérarchisé

La codification de la classification hiérarchique décimale permettra d'ordonner les uns par rapport aux autres les items dans la table QUESTIONS (voir figure 20).

Code_	Intitulé	Réponse_attend	Type_	Nature
1	LA FAMILLE	0		S
11	Nombre de personnes à reloger --	2	M	GM
111	Nombre d'adultes ----------	2	M	S
112	Nombre d'enfants ----------	2	M	GM
1121	nombre d'enfants < 3 ans	1	M	S
1122	nombre d'enfants entre 3 et 18	1	M	S
1123	nombres d'enfants majeurs	1	M	S
12	Présence d'animaux ? (Oui/Non)	1	L	GM
121	nombre de chiens	1	M	S
122	nombre de chats	1	M	S
123	autres animaux (nombre)	1	M	S

Figure 20 – Questions hiérarchisées

Cette organisation, mise effectivement en œuvre dans une application qui sera présentée au chapitre suivant, requiert un "équipement" accompagnant la seule énumération des questions :

- Tout d'abord, des informations dans la table QUESTIONS complètent les codes et intitulés des questions; ces informations sont décrites sur la figure 21. La rubrique "Type de la question" suggère que les réponses pourront avoir des types de valeurs différents ; ce point peut en effet être traité de la manière la plus simple si on suppose que la réponse, qu'elle soit de type logique, sous forme d'un choix ou d'une quantité, peut toujours être fournie par une valeur numérique. Les deux autres informations ("Réponse_attendue_?" et "Nature") permettent de gérer les titres, les questions conditionnées, à réponses homogènes ou non, exclusives les unes des autres ou multiples. La figure 20 montre ces paramètres complémentaires.

Structure de la table QUESTIONS :

- Id_question : identifiant (numérotation automatique)
- Code_question : code décimal hiérarchisé
- Intitulé : texte de la question (ou du titre de groupe de questions)
- Réponse_attendue ? : 0 → non 1 → oui, facultative 2 → oui, obligatoire
- Nature :
  S : simple
  GE : introduit un groupe de questions homogènes à réponses exclusives (une seule réponse doit être fournie)
  GM : introduit un groupe de questions à réponses multiples possibles

Réponse attendue?	Nature : S	Nature : GE ou GM
0	Titre de groupe comportant : - des titres uniquement - ou des rubriques hétérogènes	Titre de groupe de questions (homogènes si GE)
1 ou 2	Question simple (réponse comprise entre valeurs mini et maxi)	Question introduisant un groupe de sous-questions (homogènes si GE) auxquelles on ne répondra que si on a répondu à cette question par une valeur > 0

- Type :
  L : Logique (0 ou 1)
  O : Option (0 → choix_autre_que_les_suivants 1 → choix_1 2 → choix_2 ...)
  M : Montant (ou quantité)
- Valeurs par défaut, minimum et maximum pour la réponse attendue

Figure 21 - Les rubriques de la table QUESTIONS

- Une requête, s'appuyant sur la valeur du code de la question, permet d'accéder aux questions conditionnées par une question générique et de traiter les informations les concernant (par exemple, vérification de la vraisemblance d'ensemble des réponses saisies pour un groupe); un algorithme formé d'une répétitive ("Pour chaque question ...") et incluant cette requête permet de traiter l'ensemble des questions emboîtées.
- Pour pouvoir gérer l'évolution des critères dans le temps, et notamment le rajout de questions s'intercalant entre d'autres, une règle de création des codes des nouvelles questions a été édictée pour permettre de situer correctement la nouvelle question : la question se plaçant après celle dont le code est "c" aura pour code "cX5" (X jouant le rôle d'indicateur de "suite", la valeur 5 permettant ensuite d'intercaler par exemple "cX3" entre "c" et "cX5") : voir l'exemple figure 22. Un algorithme de ré-affectation de codes décimaux consécutifs et de réalisation des indentations des textes des items selon leur place dans la hiérarchie a été écrit : on obtiendra le code "122" pour la question "nombre de cochons d'Inde", et "123" pour "nombre de chats".

Code	Intitulé	Réponse attend	Type	Nature
12	Présence d'animaux ? (Oui/Non)	1	L	GM
121	nombre de chiens	1	M	S
121X5	nombre de cochons d'inde	1	M	S
122	nombre de chats	1	M	S
123	autres animaux (nombre)	1	M	S

Figure 22 – Cochons d'Inde entre chiens et chats

- Enfin, pour accéder aux données saisies et aux synthèses, une interface paramétrable pour effectuer toute sélection de questionnaires - sur autant de critères que l'on souhaite – doit être prévue (nous en montrerons une au paragraphe 6.3), ainsi qu'une requête simple donnant en une seule fois les résultats statistiques pour l'ensemble des questions.

Ainsi, l'organisation qui vient d'être présentée réunit l'ensemble des avantages recherchés, à savoir:

- une structure indépendante du nombre et des types de questions (une seule table, au lieu d'un ensemble de tables spécialisées par question), d'où une aptitude à l'évolution du questionnaire dans le temps,
- la prise en compte des questions multiples,
- celle des questions "mère-filles",
- enfin, la représentation d'une taxonomie, puisque les enregistrements de la table QUESTIONS peuvent avoir le statut de "titre" et que l'on peut hiérarchiser ces titres entre eux, comme par exemple :

```
1   Le logement ---------------------------------------- (titre)
11      Situation ------------------------------------- (titre)
111         ville (1) périphérie (2) rural (3) ----- (question de type O_Option)
112         immeuble ? ------------------------------ (question de type L_Logique)
1121            étage ? ------------------------- (question de type M_Montant)
12      Caractéristiques---------- (titre)
13      Conditions financières------- (titre)
```

6. Applications

La pratique fréquente du conseil pour l'approche et le traitement des problèmes quantitatifs nous a permis de faire la constatation suivante : Souvent, celui qui vient nous consulter a déjà réalisé un stockage des données avant d'avoir intégré toutes les dimensions du problème posé ou sans avoir tenu compte des règles de normalisation que présupposent l'usage des langages de requête (requêtes en SQL des SGBD ou, par exemple, commande "Rapport de tableau croisé dynamique" du tableur EXCEL). Il s'adresse alors à l'informaticien ou au statisticien pour demander un "réajustement technique spécialisé" qui lui permette de franchir une difficulté qu'il n'avait pas prévue ... L'expérience prouve que, le plus souvent, le meilleur conseil consiste alors à reprendre le problème sous un angle méthodologique général ("réajustement méthodologique" portant sur la précision de la problématique, la réorganisation des concepts entre eux, ...) pour résoudre ensuite la difficulté avec des moyens tout à fait classiques.

C'est pour éviter de telles dérives que, outre les actions de formation initiale ou continue, s'est développée l'orientation vers la réalisation de systèmes et procédés généraux. Tous s'appuient sur l'ensemble des principes qui viennent d'être présentés et discutés, avec, notamment, recours à des degrés de paramétrage élevés.

Nous présentons successivement une application destinée à l'expérimentation en biologie, une réalisation pour la tenue d'un panel, et une autre, de type administratif, permettant de gérer des ensembles de questions hiérarchisées et conditionnées.

6.1.L'expérimentation

A partir de 1995, nous avons développé un système général pour stocker et explorer les données des expériences conduites au laboratoire de biotechnologies végétales de l'ISAB. Son schéma général de principe est celui indiqué sur la figure 23

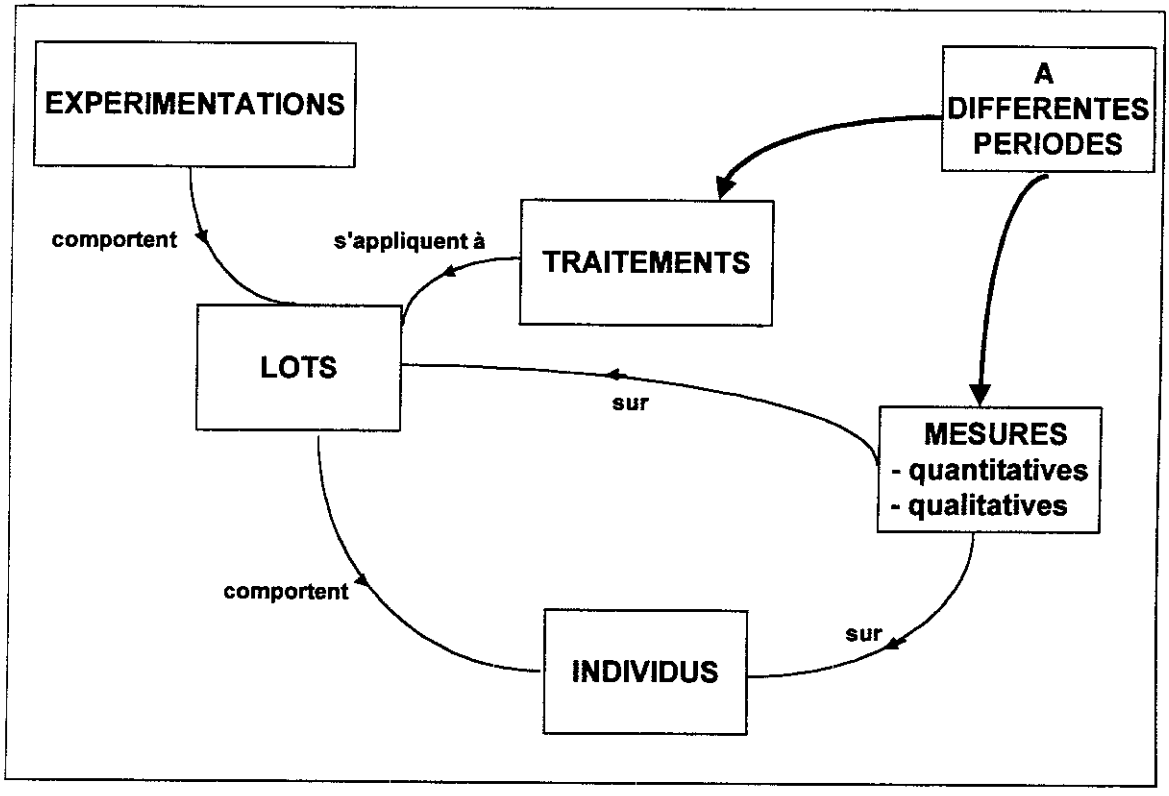


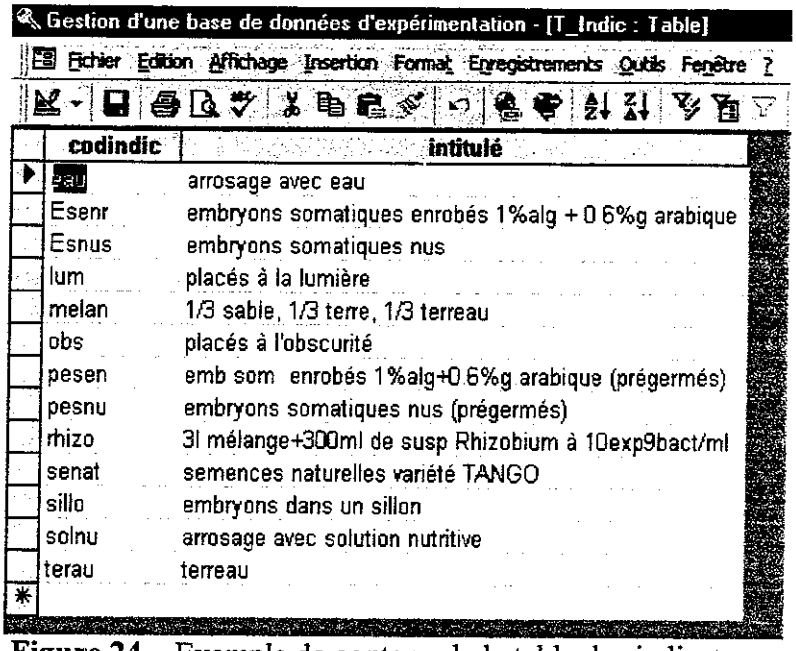
Figure 23 – Schéma général de principe du système pour les expérimentations

Ses caractéristiques et fonctionnalités principales sont les suivantes :

- Un traitement appliqué sur un lot au cours d'une période peut être décrit, dans le cas le plus général, par un ensemble d'indicateurs auxquels on peut associer une quantité (une dose) ; ces indicateurs sont gérés dans une table mise à jour par les expérimentateurs (voir figure 24).
- Possibilité de décrire des mesures, en nombre quelconque, effectuées à différentes périodes sur les lots et/ou sur les individus ; ces mesures sont de types divers : numériques – entiers ou réels -, alphanumériques, ou logiques. Les caractères mesurés sont eux aussi décrits par avance dans une table avec, selon leur type, leurs valeurs minimum et maximum ou leurs valeurs possibles.
- Présence de requêtes générales, comparables à celles évoquées pour le dépouillement de l'enquête (paragraphe 4.1), qui, pour chaque type de caractères mesurés, fournissent tous les résultats appropriés : sommes, moyennes, effectifs, fréquences,... en tenant compte des valeurs manquantes. Ces requêtes ne sont pas simples à écrire compte tenu des diverses possibilités conditionnant les résultats (obtention par lot, ou par traitement, ou par période,...) et de la gestion des divers types de variables et valeurs. L'intérêt d'en disposer réside dans le fait qu'elles permettent à l'utilisateur, dès les données saisies, de lancer sous

forme de "moulinette" des explorations systématiques fournissant une première approche des résultats expérimentaux.

- Associés à ces requêtes, création automatisée de graphiques présentant, de façon systématique pour l'ensemble des caractères mesurés (à l'imprimante) ou à la demande pour l'un de ces caractères (à l'écran), les résultats exploratoires évoqués ci-dessus.



codindic	intitulé
eau	arrosage avec eau
Esenr	embryons somatiques enrobés 1%alg + 0 6%g arabique
Esnus	embryons somatiques nus
lum	placés à la lumière
melan	1/3 sable, 1/3 terre, 1/3 terreau
obs	placés à l'obscurité
pesen	emb som enrobés 1%alg+0 6%g arabique (prégermés)
pesnu	embryons somatiques nus (prégermés)
rhizo	3l mélange+300ml de susp Rhizobium à 10exp9bact/ml
senat	semences naturelles variété TANGO
sillo	embryons dans un sillon
solnu	arrosage avec solution nutritive
terau	terreau

Figure 24 – Exemple de contenu de la table des indicateurs

La base de données réalisée comporte une quinzaine de tables reliées entre elles. Les figures 25 et 26 montrent deux paramétrages différents pour deux thèmes d'expériences différents. A titre indicatif, lors de l'étude sur la germination des embryons (figure 25), environ 100 000 valeurs mesurées ont été saisies. Un tel système, s'adaptant à des "configurations" de données si différentes, offre à l'utilisateur :

- aucun développement de tables ou de feuilles de calcul à réaliser,
- la même interface pour saisir le paramétrage de toute expérimentation (figure 27) et le même écran de saisie des valeurs observées,
- l'exploration à la demande des résultats statistiques descriptifs (sur la figure 28 : représentation graphique des résultats d'un caractère mesuré pour un lot).

Rappelons aussi que de disposer des données enregistrées dans une structure à la fois générale et conforme à la nature des concepts élémentaires de l'expérimentation permettra par la suite de composer aisément toute présentation de tableau qui serait souhaitée (et ici, même des tableaux obtenus à partir de données provenant d'expériences ayant fait l'objet de répétitions dans le temps) ou encore de réaliser tout graphique jugé adapté au problème étudié.

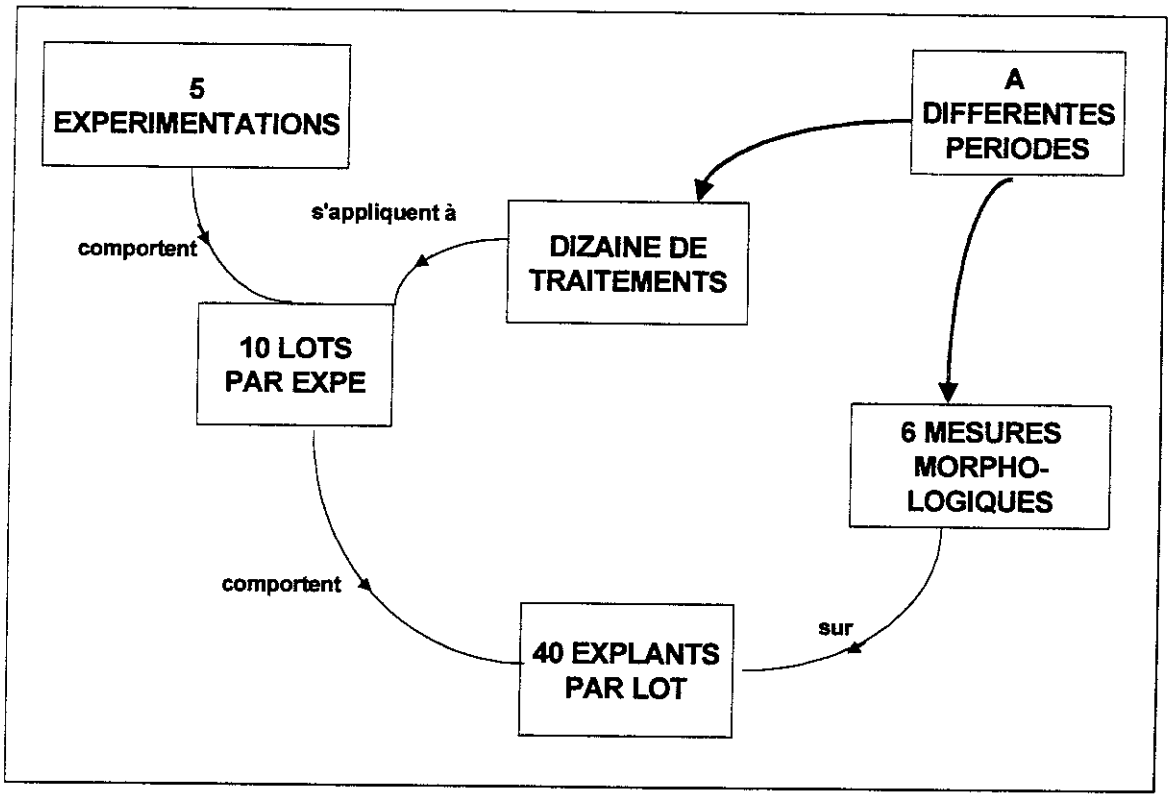


Figure 25 – Recherche de conditions favorables à la germination des embryons de *Prunus Avium*

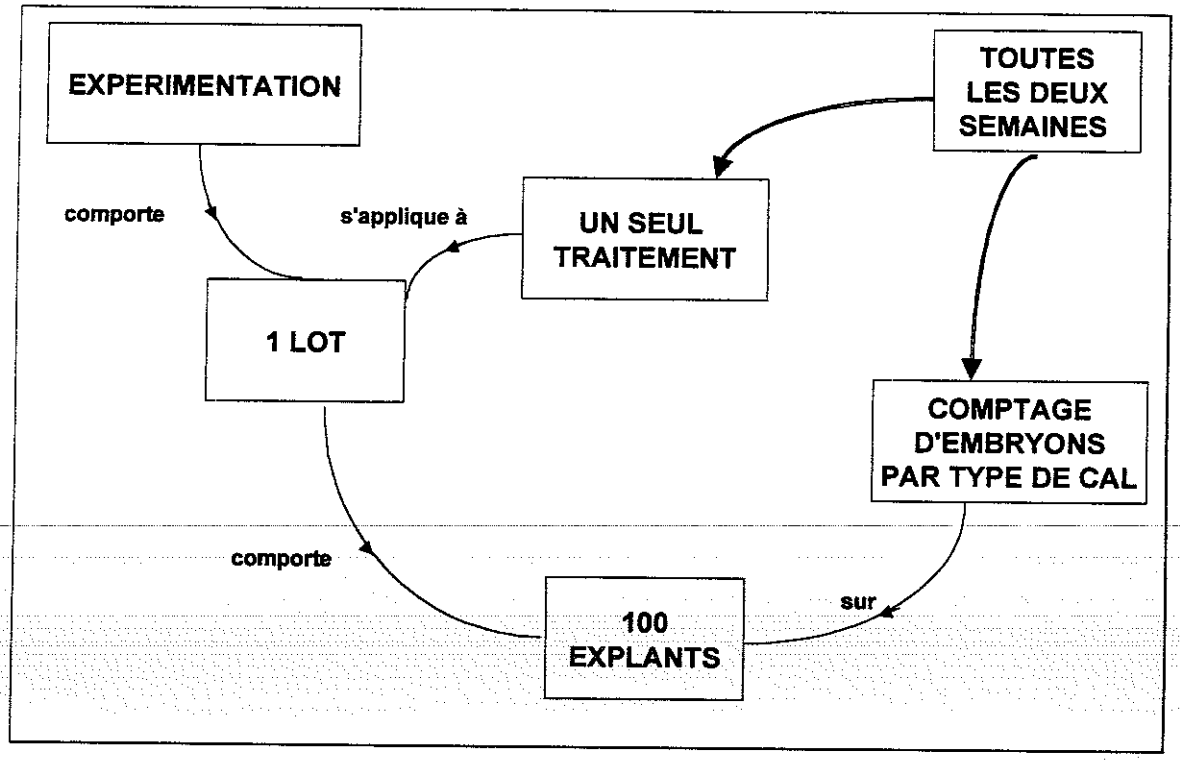


Figure 26 – Recherche de marqueurs biochimiques de l'embryogénèse chez le *Prunus Avium*

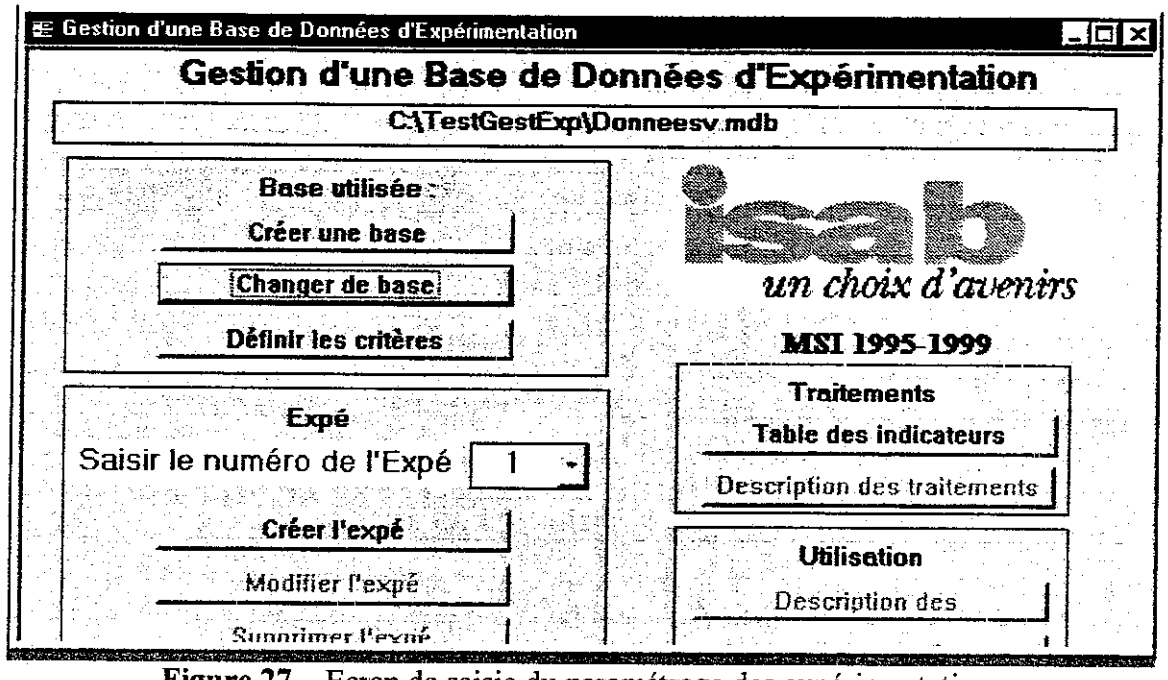


Figure 27 – Ecran de saisie du paramétrage des expérimentations

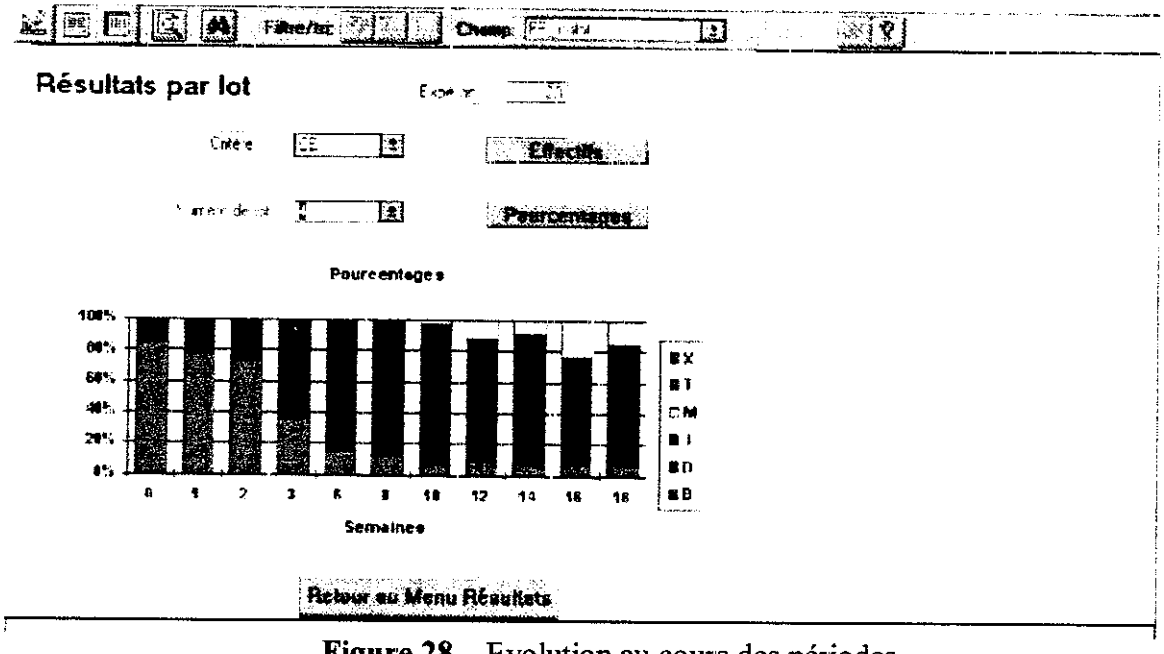


Figure 28 – Evolution au cours des périodes
des valeurs de CE (couleur de l'embryon) pour les individus du lot n°4

6.2. Les enquêtes et panels

Une structure de base de données a été préconisée il y a trois ans pour l'Assemblée Permanente des Chambres d'Agriculture qui réalise le stockage annuel et l'exploitation statistique sur plusieurs exercices des données économiques et sociales des Chambres : reposant sur certains des concepts de modélisation des panels évoqués au paragraphe 4.2, elle est aussi accompagnée de fonctionnalités de calculs de résultats synthétiques (comptages, sommations, rapports, calculs de ratios, ...) et d'un principe de codification décimale hiérarchique des rubriques (comme celui utilisé au paragraphe 5.2).

6.3. Les questions hiérarchisées dans un système de type administratif

Ce procédé a été retenu pour une application actuellement en cours de finition et de mise en exploitation :
"Tandem Immobilier" est une agence immobilière à vocation sociale située à Beauvais : Pour remplir sa mission sociale, Tandem favorise le rapprochement entre des ménages à loger et des propriétaires de logements.
Un nombre important d'informations (60 pour décrire les ménages et leur demande, 80 pour décrire les logements) sont nécessaires pour effectuer ces rapprochements tout en respectant les conditions qui permettent aux ménages et aux propriétaires d'obtenir les aides prévues par la législation. Or, l'évolution fréquente des règlements fait que rapidement certains nouveaux critères s'avèrent nécessaires tandis que d'autres deviennent inutiles. Ces deux aspects interdisaient de préconiser une description traditionnelle des caractères : cela aurait conduit à devoir modifier plusieurs fois par an les structures des tables, des masques de saisie et des traitements produisant statistiques et sélections des logements répondant aux demandes !

Aussi, nous avons décidé du paramétrage des critères pour ménages et logements dans la table QUESTIONS, les réponses étant respectivement saisies dans les tables DESCRIPTION_MENAGES et DESCRIPTION_LOGEMENTS (voir figure 29). La hiérarchisation des groupes de questions conditionnées est décrite au moyen du code décimal et des procédés complémentaires présentés au chapitre précédent (paragraphe 5.2).

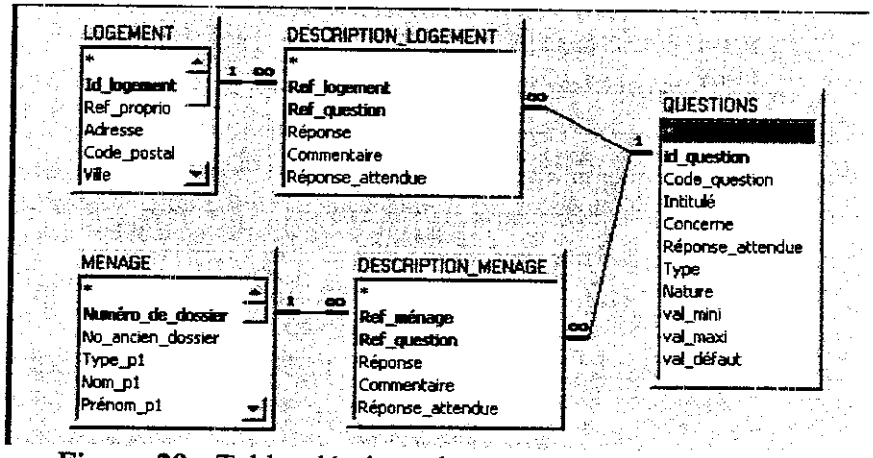


Figure 29 - Tables décrivant les ménages et les logements

La figure 30 montre la saisie des informations pour un ménage : les questions et réponses saisies apparaissent dans le sous-formulaire nommé "DESCRIPTION (du ménage)".
Outre les éditions des fiches des ménages, propriétaires et logements, deux types de résultats peuvent être obtenus :

- des sélections de ménages ou de logements entièrement paramétrables puisque incluant tous les critères souhaités présents dans la table QUESTIONS (figure 31),
- les résultats des statistiques élémentaires, obtenus en une seule requête simple (figure 32).

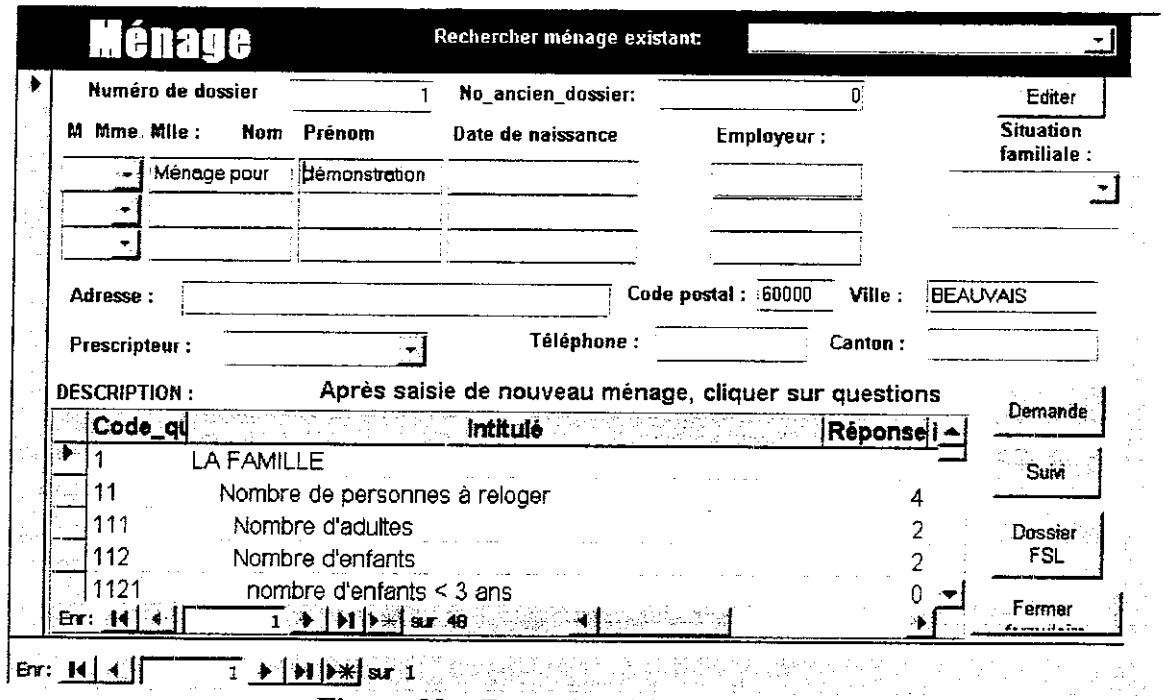


Figure 30 – Ecran de saisie des ménages

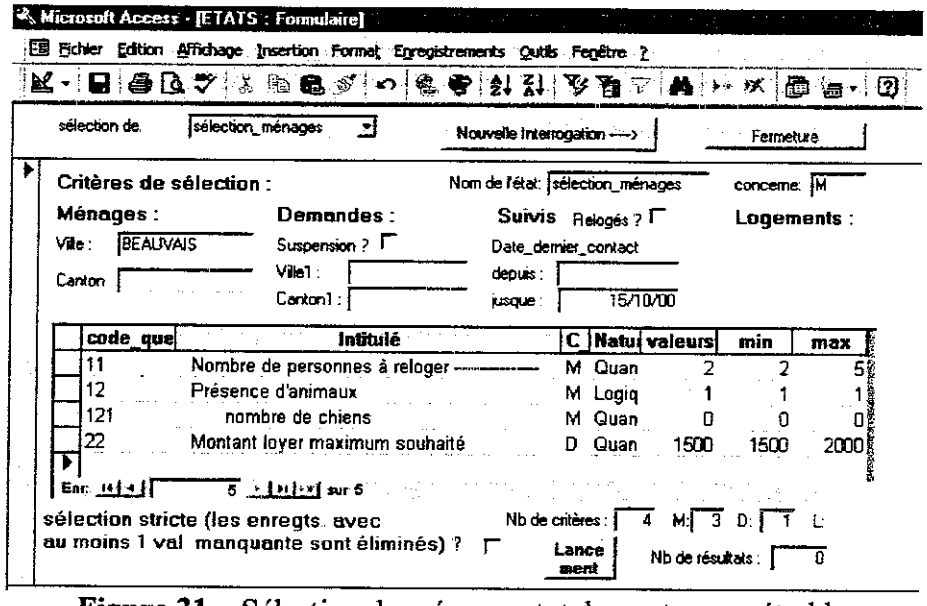


Figure 31 – Sélection de ménages, totalement paramétrable ("... et qui ont des animaux autres que des chiens")

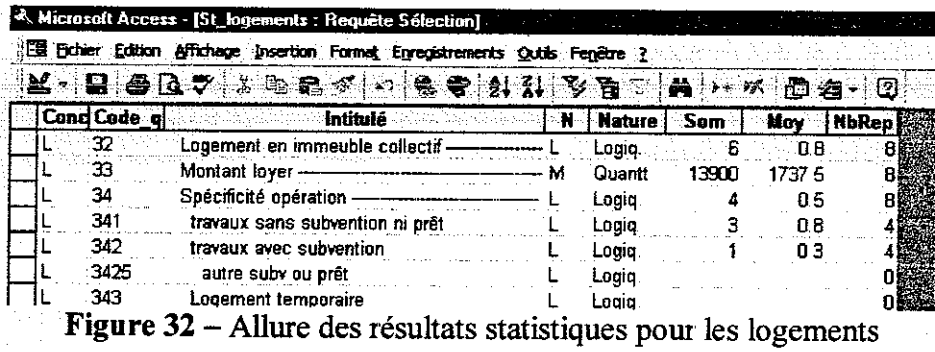


Figure 32 – Allure des résultats statistiques pour les logements

Ainsi, avec un tel outil, l'organisme en question gère toute son activité de suivi de ménages, de propriétaires et de logements; il devient aussi apte à fournir rapidement à ses prescripteurs et aux organismes financeurs toutes informations relatives à cette activité.

7. Pour conclure

Les applications présentées au chapitre précédent s'appuient sur des principes et des règles qu'il faut connaître pour la conduite et la bonne réalisation d'une étude statistique de taille significative.
Mais il est vrai que la perception de l'ensemble des interdépendances entre les divers stades et concepts évoqués (analyse du problème et modélisation conceptuelle, organisation physique du stockage des données et type d'outil de traitement, structure du tableau statistique et méthodes de traitement) n'est pas toujours accessible facilement.

Aussi, dans cette conclusion, en renommant les règles et principes, nous voulons souligner ce qui nous apparaît comme des impératifs. Et nous évoquerons aussi les potentialités, choix et risques qui se présentent alors concrètement pour l'utilisateur et le statisticien.
On remarquera que les choix se ramènent à une appréciation en termes de temps disponible, de fréquence de traitements, de coût des investissements, mais aussi d'aptitude ou de connaissance des outils, d'où, certaines fois, de la possibilité ou non d'être conseillé ou assisté.
En outre, si cet ensemble de paramètres et de conditions interdit d'être catégorique – et du coup de supprimer les choix possibles -, nous dirons qu'il est important de considérer dans une seule vision l'ensemble des options qui se présentent, en les rapprochant des contraintes spécifiques rencontrées et des impératifs à respecter.

7.1. Méthodologie et transdisciplinarité

Tout chercheur en conviendra : indispensables, elles conduisent d'abord à construire un modèle d'analyse adapté au champ et aux questions du problème abordé. La structuration correcte des concepts, de leurs dimensions, composantes et indicateurs en dépend.
De cette structuration découle celle des données utiles à l'étude. Si cette modélisation est parfois difficile, c'est que justement le problème abordé mérite d'être clarifié, précisé. Le recours à la modélisation "Entités/Associations" contribue toujours significativement à cette clarification ; elle devrait être systématique pour l'approche des problèmes complexes.
Ici, si le besoin s'en fait sentir, une étape de validation par un tiers du respect des syntaxes des langages utilisés ("Entités/Associations" et "Relationnel") ne doit pas être sautée : une formulation incorrecte compromettrait la bonne gestion des données.

7.2. Modèle ou méta-modèle ? mais organisation découlant du problème réel !

Choisir entre modèle et méta-modèle, ..., nous devrions dire, choisir entre modèle strictement et seulement adapté à un instant donné et modèle doté de caractéristiques de généralisation ; ces caractéristiques, réalisées par des procédés de paramétrage, assurent unicité de l'outil pour diverses utilisations et aptitude à intégrer les évolutions, d'où une pérennité certaine. Mais la généralisation s'accompagne de complexité ; il faut donc auparavant mesurer la faisabilité et l'opportunité d'une telle orientation en prenant en compte les critères évoqués (durées, coûts, compétences, ...).
Une fois le choix effectué, la seule manière d'être assuré de pouvoir correctement et aisément construire les tableaux nécessaires et de calculer les résultats souhaités est de disposer d'une structure conservant les données brutes repérées dans l'analyse du problème tel qu'il se présente dans la réalité.

Il faut en conséquence s'interdire de saisir les données collectées sous la forme requise par le premier traitement que l'on prévoit de faire : car la "remontée" aux diverses entités ou unités expérimentales, qui peut s'avérer utile pour la suite de l'étude, est parfois périlleuse sinon impossible (et ce, quel que soit l'outil de traitement employé).

7.3. Logiciel spécifique ou progiciel ? SGBD ou tableur ?

Les fonctionnalités qui doivent être appréciées sont celles relatives à la gestion des données et à la réalisation des traitements (celles de la sécurité des informations, et celles concernant la souplesse d'emploi et l'ergonomie), comparées à l'investissement que l'on consent.

Les particularités relatives à l'organisation de l'activité et aux compétences présentes sous-tendent ces choix : si un seul utilisateur peut se satisfaire d'un outil rustique, le travail réparti au sein d'une équipe commandera des systèmes plus élaborés.

Un usage combiné d'un applicatif spécifique et d'un progiciel est possible, une fois étudié l'interface entre les deux logiciels.

D'autre part, lorsqu'on élabore le schéma conceptuel des données d'une application (comme celui de la figure 18 au paragraphe 4.2), on trouvera intéressant d'avoir recours à un logiciel d'aide à la conception des systèmes d'information. Cet outil permet en effet de tenir au propre le graphique représentant ce schéma au fur et à mesure de l'ajout, de la suppression et de l'organisation entre eux des concepts retenus.

Une fois le schéma conceptuel terminé, le même logiciel permettra de créer automatiquement une base de données destinée à être exploitée par tel ou tel SGBD, selon le choix de l'utilisateur. Ainsi se concrétise le parcours recommandé : analyse et conception du modèle, structuration de la base de données en découlant, emploi d'un SGBD pour saisir, stocker et utiliser les données (notamment au moyen de requêtes pour les extractions ou, déjà, pour la réalisation de calculs).

Aussi, tout en ayant conscience que "le meilleur outil est peut-être celui dont on a l'habitude de se servir" et qu'une formation minimale est indispensable pour la compréhension des principes du fonctionnement d'une autre catégorie de logiciel, nous préconisons pourtant résolument d'utiliser un SGBD pour gérer les données (et de n'avoir qu'un seul lieu de stockage où s'effectuent toutes les opérations de leur mise à jour !). Et, avec la possibilité de passer des données de l'un à l'autre, on s'oriente vers un usage combiné des divers types d'outils selon leur véritable destination : SGBD, puis tableur ou logiciel statistique spécifique.

REFERENCES

[ASU96] Groupe "Logiciels "de l'ASU.- Enquête 1996 : Utilisation des logiciels statistiques.- *in* Revue Modulad, n° 21, mai 1998 - pp. 1-62.

[CERCL] Cercles d'Excel'ense- Internet : <http://www.cisia.com>

[CHR97] CHRISMENI C., LUGUET J., PUJOLLE G., ZURFLUH G.- Bases de données relationnelles.- *in* Techniques de l'Ingénieur, Paris, vol H2 - 2038, 1997 - 14p.

[COL87] COLLONGUES A., HUGUES J., LAROCHE B.- Merise, Méthode de conception.- Dunod Informatique, 2ème éd., 1987 - 240 p.

[CXP] Centre d'Expérimentation des Progiciels (CXP)- Internet : < http://www.cxp.fr>

[GAL84] GALACSI- Les systèmes d'information : analyse et conception.- Dunod Informatique, 1984 - 200 p.

[GRE98] GRENIER E., VAILLÉ J., CHAPELAIN K., REY J.F.- L'enseignement de la Statistique dans une école d'ingénieurs.- XXXèmes journées de la Société Française de Statistique, vol. II, 1998 - pp. 279-281.

[MEL85] MELESE J.- Approche systémique des organisations : vers l'entreprise à complexité humaine.- Hommes et Techniques, 1985 - 157 p.

[QUI95] QUIVY R., VAN CAMPENHOUDT L.- Manuel de Recherche en Sciences Sociales.- Nathan, 2ème éd., 1995 – 287 p.

[STE99] STÉPHAN V., LECHEVALLIER Y., HÉBRAIL G.- Statistique et Bases de Données.- *in* Actes de la journée "Statistique et Bases de Données" de la Société Française de Statistique, 24 juin 1999 – 49 p.