DISCUSSION SUR LA REALISATION ET L'APPLICATION D'UN LOGICIEL DE TRAITEMENT D'ENQUETES

Agnès Guillaume et Philippe Hamelin
(UR 205, Dynamique des Populations et Culture, Enjeux et Maîtrise de l'Espace)

Museu Emilio Goeldi
C.P. 075 ORSTOM
66000 Belém, PA (Brésil)

RESUME - La réalisation d'un logiciel de saisie de données semble une tâche relativement simple quand l'on possède quelques connaissances informatiques, une expérience de l'organisation des données et que l'on a identifié les besoins des utilisateurs. Dans ces conditions, en dépassant le budget-temps prévu pour la phase de mise au point, on parvient à diffuser à quelques chercheurs un produit qu'ils utilisent avec satisfaction. Mais l'étape suivante, la transformation du logiciel en un produit de niveau commercial, est trop semé d'embûches pour qu'un chercheur puisse la franchir seul, sans le soutien de services compétents pour aider à la normalisation du produit, en assurer la diffusion et la maintenance.

Nous participions à la réalisation, depuis quelques années, de traitement d'enquêtes à l'aide de logiciels implantés sur le centre de calcul d'Orsay (CIRCE). La maîtrise de l'analyse statistique proprement dite ne posait plus guère de problèmes ; par contre, l'étape saisie et dépuration des données ressemblait trop souvent à un casse-tête chinois.

Cette étape exigeait un travail long, souvent rébarbatif et peu créatif. Cette phase du travail avait souvent pour conséquence que l'analyse de certaines enquêtes n'était que très partielle et, dans les cas extrêmes, le seul traitement réalisé était le chargement des données sur bande magnétique et l'archivage de celles-ci.

Ce phénomène était aggravé du fait que certains chercheurs, croyant en la magie du mot ordinateur, avaient un peu perdu de vue les problèmes de la qualité des données et des supports sur lesquels elles étaient recueillies. L'ordinateur était la boîte noire où l'on jetait un paquet de cartes à l'entrée pour récupérer un rapport à la sortie : nous avions enfin trouvé notre pierre philosophale!

Ce mystère s'entretenait d'autant plus facilement qu'avec l'apparition de la statistique descriptive multivariée (analyse des

correspondances, analyses factorielles, etc.), une confusion s'est très vite faite entre la capacité de ces outils à analyser des données hétérogènes (qualitatives, quantitatives, ordinales, etc.) et à analyser n'importe quoi.

Même si le résultat à la sortie de la machine n'était pas probant, l'appel au joker (recodage, lissage, générateur de données manquantes) permettait toujours au bout d'un certain nombre de passages dans la machine d'obtenir un résultat, mais quel résultat ?

Il faut dire que ce système arrangeait tout le monde : le chercheur à qui il évitait une remise en cause de son travail de recueil (de toutes façons les résultats seraient bien plus valorisés par la méthodologie employée que par leur valeur intrinsèque) ; l'intermédiaire informaticien/statisticien voyait grandir son prestige et devenait un homme indispensable et les institutions de tutelle montraient que leurs troupes faisaient preuve de modernisme.

Nous étions dans la situation paradoxale où l'introduction d'un puissant outil technologique aboutissait à l'apprauvissement des résultats. Cet état de fait se vérifiait surtout dans les sciences où la méthodologie ne possédait pas d'outils mathématiques ou statistiques développés.

L'arrivée en force des micro-ordinateurs, au début des années 80, allait permettre cette démystification de l'ordinateur à travers sa démocratisation. La réalisation d'un logiciel de saisie, contrôle et traitement élémentaire des données nous apparaissait comme une façon de valoriser ce rapprochement de l'ordinateur et du chercheur.

Nous nous sommes cantonnés à la réalisation de traitements statistiques élémentaires, car si la saisie et la vérification de données ne sont que l'automatisation de tâches répétitives ne demandant pas de connaissances théoriques, le fait de savoir positionner cinq paramètres pour lancer une analyse de correspondances ne garantit en rien la maîtrise des présupposés théoriques, qui seuls permettent une interprétation scientifique correcte des tableaux et des graphiques fournis par l'ordinateur. Sans cette connaissance préalable de la statistique, l'interprétation est un acte de divination dans le marc de café.

La deuxième grande préoccupation était le temps : le temps très long entre la passation de l'enquête sur le terrain et la sortie des premiers résultats. Une durée de trois ou quatre ans entre les deux était considérée comme moyenne et les résultats devenaient parfois obsolètes avant publication. Souvent aussi les chercheurs attendaient leur retour en France pour traiter leur enquête, ce qui interdisait le plus souvent tout contrôle *à posteriori* des doutes de toutes sortes qui pouvaient surgir à la vue de certains résultats.

1. LE PROJET

A partir de l'analyse ci-dessus, et à la demande de nos collègues démographes, ont été déterminées les fonctionnalités de ce que devraient être un logiciel informatique capable d'améliorer la qualité des données en rapprochant l'outil informatique du chercheur sur le terrain, de raccourcir cette étape rébarbative entre enquête et analyse, et sous la condition que l'investissement purement informatique de l'utilisateur soit minimum.

1.1. Fonctionnalité générale du logiciel

La facilité d'utilisation, aucune connaissance préalable en informatique ne devrait être requise, ce qui entraînera la réalisation d'un logiciel fermé où l'utilisateur sera guidé pas à pas par une série de menus. Une coupure de courant, où la ballade d'un chat sur le clavier, ne devrait se traduire que par la perte du dernier enregistrement au maximum.

1.2. La structure

Il a été choisi une structure à deux fichiers, un fichier contenant les descripteurs des variables (le dictionnaire) et un fichier contenant les données proprement dites. Cette structure est celle des grands progiciels statistiques (OSIRIS, SPSS, SAS) mis en oeuvre au CIRCE. Les chercheurs en sciences humaines traitant leurs données majoritairement sur ce centre de calcul ne devraient pas, de ce fait, se sentir dépaysés.

Une réflexion avait été menée sur la structure hiérarchique des fichiers (économie de place en mémoire de masse), mais cette solution a été abandonnée parce que les difficultés à affronter n'étaient pas à la mesure des résultats espérés. Il faut de toutes façons revenir à une structure rectangulaire pour les analyses : le choix, pour les données très hiérarchisées, a été de les éclater en plusieurs fichiers.

2. LE MATERIEL

Le choix du micro-ordinateur Goupil 3, muni du système d'exploitation Flex 9, était évident à cette époque (décembre 1983), puisque :

- c'était l'unique matériel dont nous disposions ;
- c'était aussi le micro-ordinateur retenu pour la diffusion à l'ORSTOM.

Ce micro-ordinateur était équipé d'un double lecteur de disquettes 8", d'une capacité de un million d'octets. Le choix de

l'interpréteur SBasic allait de soi à cette époque. L'autre choix possible d'un développement en assembleur 6809, malgré la simplicité de cet assembleur, aurait eu pour conséquence une multiplication rédhibitoire du temps de développement.

Le SBasic, par rapport au GWBasic présent sur les machines MS-DOS, présente de nombreux avantages, comme la gestion simple de l'écran et des routines d'erreur ou la présence de macro-fonctions permettant de générer du code (interprétable à l'intérieur du programme lui-même), qui donnait la possibilité de développer facilement des microlangages pour certaines parties du logiciel.

Enfin, la possibilité d'utiliser sous l'interpréteur Basic, à n'importe quel moment, les commandes du système, autorisait la réalisation d'un logiciel totalement fermé.

3. LE LOGICIEL

Il se décompose en trois grands modules et quelques programmes utilitaires. Chaque module est un assemblage de petits programmes (au maximum 250 lignes), qui sont chaînés les uns aux autres à travers une série de menus. Cela était nécessaire car nous ne disposions, après le chargement de l'interpréteur Basic, de guère plus de 30K, pour les programmes et les données.

Les fichiers sont du type à accès direct. Les données numériques sont enregistrées en binaire sur deux octets et les données alphabétiques en caractères. En réalité les données numériques et alphabétiques sont situées sur deux fichiers parallèles, mais cela reste totalement transparent pour l'utilisateur.

3.1. Le module dictionnaire

Le module dictionnaire est composé de quatre programmes :

- le premier assure la création du fichier dictionnaire et du fichier de données, ainsi que leur protection contre toute destruction intempestive ;
- le deuxième enregistre, à travers une saisie contrôlée, les descripteurs de chaque variable. Ces descripteurs de variables sont au nombre de huit : le nom, le nom abrégé, le type, la longueur, la valeur minimale, la valeur maximale, le nombre de valeurs d'exclusions (valeurs en dehors des bornes extrêmes qui servent à repérer, par exemple pour les âges, les gens dont on ne connaît pas l'âge, mais dont on sait s'ils sont adultes ou enfants), et enfin la valeur de la *"sans réponse"* ;
- le troisième permettra la visualisation, l'impression et éventuellement la correction des donnés enregistrées avec le deuxième programme ;

- le quatrième assurera, à l'aide de données enregistrées avec le deuxième programme, la génération des paramètres du masque de saisie. Il calculera, en fonction du nombre d'individus prévus en saisie, la taille à réserver sur la disquette et effectuera cette réservation. En cas de fichier dépassant la taille d'une disquette il fera les réservations nécessaires (une procédure est prévue au niveau de la saisie pour pouvoir augmenter la capacité disponible).

3.2. Le module saisie

Trois programmes composent le module de saisie des données :

- le programme de saisie qui assure l'affichage de la grille à l'écran, la gestion du curseur, le contrôle de la longueur de la variable et si celle-ci se situe bien entre les bornes définies ou correspond à une valeur d'exception. La touche *"retour chariot"* correspond au code sans réponse quand une valeur a été définie lors de la création du dictionnaire : dans le cas où aucune valeur n'a été indiquée, l'opérateur doit entrer obligatoirement une valeur. Cela est pratique : dans le cas où une partie des réponses est facultative, il est très conseillé alors de mettre une valeur obligatoire en en-tête de chaque nouveau paragraphe ;

- un programme permettant l'impression des enregistrements sélectionnés par les valeurs d'une borne inférieure et d'une borne supérieure, soit à l'écran, soit sur l'imprimante ;

- un programme de correction/vérification des données qui permet, soit de visualiser séquentiellement chaque enregistrement en avant ou en arrière, soit de se déplacer aléatoirement en fournissant un numéro d'enregistrement. Les valeurs entrées lors de la correction sont, bien entendu, vérifiées de la même façon que lors de la saisie. En réalité ce programme est bien plus performant que le programme de saisie.

Ce module pourrait être simplifié en fusionnant le programme saisie et le programme correction en un seul fichier, ce qui serait relativement simple et il suffirait de quelques ajouts mineurs au programme de correction pour qu'il puisse aussi assurer la saisie.

3.3. Le module traitement

Comme il a été dit par ailleurs, la vocation de ce module n'était pas le traitement scientifique des données mais plutôt un ensemble permettant la sortie rapide des résultats et de vérifier ainsi la cohérence, aussi bien interne qu'externe, des données.

Si la cohérence interne ne demande généralement qu'un retour au questionnaire pour être validée, la cohérence externe, mise en évidence par des tabulations simples ou croisées, demande

bien souvent un retour sur le terrain pour être décidée, d'où l'importance d'une disponibilité rapide des résultats.

Le programme de contrôle de cohérence est prévu pour tester la cohérence des réponses à deux ou plusieurs variables. Pour formuler les équations à tester, l'opérateur a à sa disposition un micro-langage qui contient comme alphabet soit des constantes numériques, soit les variables numériques décrites dans le dictionnaire, des opérateurs logiques (*ET, OU*) et des opérateurs de relation (=, <, >, <=, >=, <>). Cela est complété par le choix laissé à l'opérateur d'extraire la sous-population définie par l'équation ou son complément.

Deux modes de fonctionnement sont proposés :

- soit *interactif*. L'opérateur rentre l'équation et obtient le résultat de suite à l'écran ;
- soit *en temps différé*. L'opérateur rentre une série d'équations et lance l'exécution pratique lorsque les fichiers sont importants.

L'utilisation de ce programme permet l'extraction de sous-populations.

Le programme de tris à plat est tout simple : il suffit de fournir comme paramètre l'intervalle des variables que l'on désire traiter.

Le programme de tabulations croisées permet de traiter simultanément une, deux ou trois variables. Il existe une possibilité de filtrer les individus statistiques. La construction des filtres à l'aide du même micro-langage que pour les contrôles de cohérence. Il est aussi possible de redéfinir le minimum, le maximum et les valeurs d'exception de chacune des variables : dans ce cas les valeurs apparaissant en dehors des nouvelles limites rejetteront l'individu dans la catégorie rebut, qui ne sera pas prise en compte lors du calcul des statistiques.

Enfin, un petit programme permet de transformer les fichiers de données en binaire et à accès direct, en fichiers séquentiels EBCDIC, en vue de leur transfert et/ou de leur utilisation avec d'autres progiciels de traitements statistiques.

En plus des utilitaires de sauvegarde et de formatage de disquettes, il a été réalisé au Togo, par l'équipe Levy/Pilon, d'autres programmes, en particulier, un de recodage qui faisait défaut.

4. LA REALISATION

Le projet était peu ambitieux au départ, la durée prévue pour sa réalisation était de deux mois. Il est vrai que fin janvier une première version du logiciel était mise en test, le module traitement

ne comportant alors qu'un programme. Mais au minimum deux mois et demi ont été nécessaires pour les tests et la mise au point.

Nous avions sous-estimé la durée de la phase test et mise au point, qui est certainement la plus longue. Nous estimons la phase définition du projet aux seuls quinze jours de décembre pendant lesquels nous avons défini les fonctionnalités du projet et à un mois l'écriture des programmes.

L'impasse faite sur l'analyse organique a eu un impact direct sur une écriture embrouillée des programmes, qui sont parfois d'une maintenance difficile.

Il manque aussi une documentation des programmes ; par contre le manuel de l'utilisateur a été rédigé.

Nous nous sommes aperçus alors que le coût de développement d'un tel logiciel est très élevé, l'estimation minimum est de 200 000 francs (main d'oeuvre, matériel, locaux), ce coût aurait certainement doublé pour mettre le produit à une norme commerciale. heureusement qu'il n'existait pas de produit semblable disponible sur le marché!

5. LES APPLICATIONS

L'enquête mortalité prénatale en Guadeloupe (enquête conjointe ORSTOM, DASS - Guadeloupe et INSERM) a été la première application de ce logiciel. La demande avait été formulée par J.P. Guengant (UR 709), qui souhaitait pouvoir exploiter cette enquête sur place afin d'en contrôler les résultats au fur et à mesure de sa réalisation.

L'enquête avait pour originalités :

- sa durée dans le temps, deux ans ;
- sa longueur et la complexité du questionnaire, 343 variables.

L'utilisation du logiciel s'est faite sans problème par des personnes sans aucune formation informatique. Une secrétaire de la DASS, qui n'avait encore jamais pratiqué l'informatique, a assuré la saisie et la correction des données. Un des médecins enquêteurs a présenté des tableaux de résultats de l'année 1984 lors d'un séminaire en février 1985. Il faut souligner que ce médecin utilisait pour la première fois l'informatique.

Une version disque dur du logiciel fût réalisée fin 1984 pour le traitement d'une enquête pluridisciplinaire au Nord-Togo. Une panne de disque dur seulement résolue à Paris fût le problème le plus important affronté par cette équipe.

Cette enquête se décomposait en sept fichiers, qui représentaient environ un total de quatre millions d'octets à saisir. Les chercheurs ont apprécié la facilité du micro-langage pour les demandes

de contrôle de cohérence, mais ont regretté sa lenteur d'exécution sur gros fichiers.

Par ailleurs ils ont fait développer des compléments par des informaticiens locaux. Ces modules complémentaires ont bien fonctionné, ce qui pourrait signifier que notre programmation n'était pas aussi illisible que nous le croyions.

Au Togo, il a été à nouveau utilisé pour une enquête épidémiologique, en Côte d'Ivoire pour la saisie des registres de dispensaire et aussi pour d'autres saisies en Guadeloupe.

6. LE BILAN

D'un point de vue strictement comptable, nous pouvons dire que le projet a été au moins amorti par les utilisateurs, en terme de gain de temps et d'une meilleure qualité des données. Mais si l'on considère que ce projet était le ré-investissement de plusieurs années d'expériences acquises par les auteurs, c'est un échec du point de vue de la valorisation de notre savoir-faire.

Les causes sont principalement de deux ordres :

- l'arrivée sur le marché et son hégémonie rapide du standard IBM-PC, qui rendait notre logiciel incompatible avec une grande partie du nouveau parc ordinateur de l'ORSTOM. Si ce logiciel avait pu être rendu compatible IBM-PC par une conversion sous MS-DOS (travail non réalisé par manque de temps de notre part, mais aussi par manque de soutien), son utilisation aurait été assurée pour des travaux en Côte d'Ivoire, Indonésie, Brésil et Mexique ;

- le manque d'infrastructure à l'intérieur de l'ORSTOM pour maintenir et diffuser les logiciels. Les deux auteurs avaient entrepris une recherche méthodologique finalisée : il n'entrait ni dans leur compétence, ni dans leur intention de se transformer en commerciaux pour vendre leur produit. Nous nous étions adressés à l'Unité de Valorisation de la DIVA pour examiner conjointement les suites qui pourraient être données et de quelle manière, mais nos démarches n'ont pas abouti. Comme d'autres engagements allaient bientôt nous accaparer, nous décidions alors d'arrêter là notre projet : nous imaginions mal comment les deux auteurs de ce projet, l'un étant au Brésil, l'autre en Côte d'Ivoire, pourraient assurer la conversion, la diffusion et la maintenance de ce logiciel.

Si le premier point est inhérent à la dynamique du développement de l'informatique, le deuxième doit être résolu avant d'envisager le développement de tout logiciel d'intérêt général.

La création d'une structure à l'ORSTOM capable, d'une part, d'évaluer les projets soumis pour décider de l'opportunité de les développer et, d'autre part, fournir une assistance technique

durant la phase de développement ainsi que d'en assurer par la suite la diffusion et le suivi est très souhaitable.

La puissance des micro-ordinateurs et la diffusion des disques durs permettent maintenant l'utilisation de puissants logiciels de traitement de données, qui, jusqu'alors, étaient réservés aux centres serveurs : se résout ainsi le problème *"analyse des données"*, néanmoins le problème de l'acquisition des données demeure.

Avant de développer un ou plusieurs logiciels de saisie, il semble indispensable d'entamer une réflexion approfondie sur les différents types de données qui peuvent se présenter. Nous pouvons déjà en distinguer deux :

- les données fournies par les instruments de mesure, où le problème est l'automatisation complète de la chaîne ;
- celles recueillies par un observateur humain, subdivisées en deux groupes :
- les enquêtes lourdes qui recueillent une grande masse de données, dans un temps limité ;
- les panels, les enquêtes à passages répétés, les observatoires, etc., soit toutes les structures qui recueillent des informations étalées dans le temps et dont le volume moyen journalier est raisonnable.

En général, la disponibilité de l'information déjà recueillie peut être d'un grand bénéfice. Si dans le premier cas le recours à des services spécialisés semble la meilleure solution, pour des raisons d'efficacité (durée de saisie, personnel temporaire qualifié), l'introduction du micro-ordinateur dans le second cas devrait dépasser le simple stade du convertisseur, pour fournir : suivi de planning, fiche enquêteur, bordereaux du prochain passage, etc..

Le futur est peut-être le micro-ordinateur portable, qui substituera questionnaire et cahier de notes. Une réflexion doit s'engager sur cette perspective qui bousculera beaucoup de concepts en matière d'acquisition de données.

CONCLUSION

L'ORSTOM n'a ni la vocation, ni les moyens, de développer une recherche en informatique, mais doit rechercher la meilleure utilisation possible de cet outil et nous croyons qu'à l'avenir deux grandes voies, qu'il ne faut impérativement pas confondre, se dessinent :

- *le service chercheur*. L'ordinateur comme bonne à tout faire du chercheur : le chercheur s'informatise pour gagner du temps lors des tâches de routine (traitement de texte, gestion de bibliographie, consultation de bases de données, calculs, etc.) ;

- *la valorisation*. Des recherches bien menées, qui aboutissent à des méthodologies fiables ou, aussi, des méthodologies développées parallèlement à certaines recherches d'un intérêt général peuvent être valorisées à travers la réalisation d'un produit informatique.

Cette transmission, à travers de l'informatique, du *savoir-faire* du chercheur équivaut, peut-être avec un sens plus large, à la réalisation du système-expert décrit par Pascal Renaud (Cf. ORSTOM Actualités, nº17). Un système-expert à l'intention des décideurs et des spécialistes chargés de l'application, mais aussi à l'intention des chercheurs d'autres spécialités confrontés à des problèmes spécifiques.

Encore faut-il bien faire la distinction entre expert et chercheur, le premier se caractérise par un diagnostic rapide grâce au savoir-faire accumulé, tandis que le second recherche l'accumulation de nouvelles connaissances à travers l'analyse d'un phénomène-recherche. Ce sont deux métiers différents, même si le chercheur joue le rôle d'expert, de temps à autre.

Cela permettrait peut-être de clarifier les débats sur la recherche pluridisciplinaire, car bien des projets dits *"pluridisciplinaires"* ne sont bien souvent que des projets monodisciplinaires qui nécessitent l'intervention de consultants ou d'experts de divers autres champs de connaissance.

L'emploi de tels systèmes ne doit, en aucun cas, tuer la créativité du chercheur et il faut rappeler que l'informatique est l'outil et que jamais une enquête ou un recueil de données ne peut-être considéré mauvais, parce que non traitable par tel ou tel logiciel informatique : ce serait au contraire la preuve de l'inadéquation , ou de la faiblesse, de ce logiciel.