Des bases de données sur le terrain ?

- Bernard Bel -

Chercheur C.N.R.S. - C.S.H.

Il est de plus en plus commun pour les chercheurs en sciences humaines et sociales de transporter un ordinateur sur le terrain. Ceci, malgré les incertitudes du climat, des sources électriques, l'éloignement des services de maintenance et les risques de destruction ou de vol. On peut se demander si l'utilisation dans de telles conditions d'un outil aussi coûteux que fragile n'est pas un luxe inutile, d'autant plus si on l'a payé de sa poche…

A quoi bon s'encombrer d'un ordinateur s'il ne sert qu'à anticiper la saisie des données que l'on pourrait commencer en toute quiétude au retour ? Dans quelle mesure est-il rentable, au niveau de la recherche, de consacrer à cette saisie un temps qui pourrait être mieux utilisé à travailler ou à établir des relations conviviales avec les informateurs ? L'informatique impose une certaine rigueur à la saisie des données, rigueur qui est elle-même fonction des techniques de modélisation sur lesquelles sont basés les logiciels de traitement de l'information.

Un premier avantage à saisir des données sur le terrain est de pouvoir en faire un pré-traitement qui permet de corriger des erreurs de modélisation, voire de s'interroger sur la pertinence des paramètres de l'enquête. De plus, il arrive que le chercheur (surtout débutant) doive rester plusieurs semaines sur son terrain pour approfondir ses connaissances du milieu social et culturel ou attendre certains événements importants pour son enquête. Dans ce cas, le pré-traitement des données est une excellente occasion de clarifier les méthodes de recherche, et surtout d'essayer de mettre sur pied des procédures de validation des modèles par les informateurs eux-mêmes.

Des données, mais dans quel état ?

On voit de plus en plus des jeunes chercheurs revenant de " la campagne " avec un ordinateur couvert de poussière, et parfois en piètre état, mais bourré d'informations glanées au cours de leur séjour. On est pourtant déçu de la manière dont ces informations ont été emmagasinées dans la machine. Tel (le) sociologue a rédigé une sorte de journal de bord de plusieurs centaines de pages, indéchiffrable parce qu'il tient dans un seul document de texte qu'on a bien du mal à faire dérouler sur un écran. Il faut imprimer tout ce fatras et tout recommencer ou presque : des semaines de travail en perspective qui auraient pu être évitées si l'auteur avait pris la peine, avant son départ, d'étudier les styles et la représentation " plan " de MS-Word. Tel (le) anthropologue a saisi des milliers de réponses à des questionnaires, croyant que MS-Excel était une base de données. On ne peut plus naviguer dans son immense tableau ni faire de requête élaborée, et pas même l'imprimer proprement. Nous voilà partis pour une conversion en base de données ; hélas, de nombreuses cases du tableau contiennent des expressions impossibles à convertir, ce qui veut dire qu'il faudra compléter manuellement les données escamotées par la conversion… Le reproche que je ferais à ces deux personnes est d'avoir cru que les logiciels généreusement offerts par leur vendeur de Mac ou PC, voire piratés chez un collègue, suffiraient à couvrir leurs besoins. Je me ferai un peu provocateur en suggérant aux chercheurs en sciences sociales de mettre à la corbeille leur traitement de texte et de se procurer les logiciels suivants :

- un éditeur hypertexte (par exemple HotMetal Pro)
- un logiciel de base de données (par exemple FileMaker Pro)

- un tableur (par exemple MS-Excel) pour le traitement de données numériques
- un ou plusieurs logiciels spécialisés, selon la discipline : statistiques, cartographie…

L'idée subversive est la suivante : il faut toujours structurer les données quand on les entre dans un système informatique. Or, structurer n'a rien à voir avec formater des paragraphes ou des titres, ce que l'on fait - en général fort mal - avec un traitement de textes… Informellement, il y a deux types de structures : régulières ou irrégulières.
Une structure régulière est représentée par des ensembles homogènes d'objets du même type (par exemple des personnes, des lieux, des objets, des événements…). Ces ensembles formeront une base de données relationnelle.
Une structure irrégulière, par contre, est caractérisée par des liens arbitraires que l'hypertexte peut rendre à merveille - voir les pages " web " bien conçues. Hypertexte et bases de données se complètent et peuvent même communiquer entre eux dans la mesure ou l'on peut insérer un lien hypertexte dans une base de données, ou réciproquement un formulaire interactif d'interrogation de base de données dans un document hypertexte. Je n'entrerai pas dans les détails techniques des ces opérations (que je ne maîtrise pas encore suffisamment).

Pertinence, précision, généralité : le triangle infernal

Cette " flexibilité " de l'information structurée, qui est apparue tout récemment avec l'invention des markup languages, dont HTML n'est qu'un avant-goût dans le domaine du texte, et la popularisation des bases de données relationnelles, répond à un besoin bien particulier des sciences sociales.

Le chercheur doit trouver un compromis entre trois caractéristiques inconciliables de tout modèle : sa pertinence, sa précision et sa généralité. Un modèle général et précis, comme c'est le cas de nombreux modèles formels, a peu de chances d'être pertinent (relevant) à une situation donnée. Par ailleurs, un modèle soigneusement élaboré à partir de données multiples du terrain, et avec grande précision, aura difficilement une portée générale. Enfin, un modèle général décrivant avec justesse une situation concrète doit souvent se contenter de gommer les détails.

Est-ce possible ?

Le présent article est limité à une discussion, sans aucune ambition théorique, des bases de données relationnelles. La littérature méthodologique sur l'hypertexte est très abondante grâce aux conférences internationales consacrées à ce sujet ces dix dernières années. Jusqu'à une époque récente, la construction d'une base de données était un travail typique d'informaticien. Il fallait soigneusement définir la structure de la base, les champs, les types de données et les procédures de vérification de cohérence, avant même de commencer à saisir des données. Paradoxalement, cela supposait une connaissance parfaite a priori du modèle de représentation des connaissances, alors que les méthodologies les plus récentes suggèrent de faire " émerger " le modèle à partir des données, ne serait-ce qu'en ce qui concerne les questions de terminologie, les concepts et autres folk views. Cette approche pragmatique est possible aujourd'hui avec les logiciels modernes.

Choisir un logiciel de base de données

Ce choix n'est pas vraiment critique car il est facile aujourd'hui de convertir les données d'un logiciel vers un autre, y compris entre Mac et PC. Une telle conversion se réduit dans les meilleurs cas à un drag-and-drop du fichier ancien sur l'icône du logiciel nouveau. Si cela ne fonctionne pas, on peut toujours exploiter les procédures d'exportation et d'importation, en passant notamment par le format " texte". La préférence pour un non-technicien ira vers un logiciel convivial plutôt que vers un environnement exigeant des qualités de programmeur.
Au niveau du modèle, le choix actuel se fait entre les bases de données relationnelles, dont il est question

dans cet article, et le modèle orienté objet qui combine plusieurs approches et dont l'intérêt premier est de traiter des données non homogènes. L'approche orientée objet est encore une affaire de spécialistes et me paraît un peu prématurée, sauf à y consacrer un temps de formation suffisant avant le début des travaux de recherche. Au CSH nous avons opté pour FileMaker Pro de Claris, une base de données relationnelle sur Mac et PC, récemment déclaré le programme le plus " cool " de l'année par un magazine informatique américain. Il permet la mise en réseau, y compris sur Internet, et le partage des même bases entre les Macs, les PC, et même des clients Unix grâce à un récent plug-in. J'ai remarqué (sur une dizaine d'années d'utilisation) que ce logiciel était particulièrement robuste aux risques de pertes de données par plantage de l'ordinateur, défaillance des disques durs ou erreurs sur le réseau local. A remarquer aussi que la combinaison de FileMaker Pro et de l'Indian Language Kit d'Apple (distribué en Inde) permet de gérer des fichiers contenant des données dans plusieurs langues locales, avec des recherches et des tris conformes à ces alphabets. Cette combinaison utilise les standards WorldScript et Unicode qui n'existent pas encore sous Windows. Enfin, FileMaker Pro communique de manière dynamique avec d'autres logiciels comme MS-Excel à qui il peut par exemple expédier des données à afficher sous forme de graphiques.
Il est important de s'assurer que le logiciel de bases de données permet de traiter les types de données non-textuelles particulières à l'étude, par exemple des images numérisées, des sons ou des séquences vidéo QuickTime. Les concurrents intéressants de FileMaker Pro sont, à ma connaissance, 4D et Oracle. 4D dispose d'un environnement de programmation plus évolué que les scripts de FileMaker Pro, et Oracle permet quant à lui des requêtes plus élaborées. Microsoft propose aujourd'hui un logiciel de base de données utilisant, comme Oracle, le langage de requête SQL, et distribué gratuitement (ceci pour contrer " l'hégémonie " d'Oracle - 40% du marché !).

"Atomiser " les données

Atomiser les données veut dire en premier lieu identifier les ensembles " d'objets " à représenter, et à créer un fichier pour chaque ensemble. La base de données est composée de ces fichiers liés dynamiquement entre eux. Une erreur fréquente au début est de chercher à tout faire tenir dans un fichier unique, par exemple un ensemble de fiches qui contiennent le nom d'un informateur, son adresse, des détails sur son lieu d'habitation, sa famille, sa profession, et les résumés des interviews. C'est ce genre de fichier que les chercheurs ont tendance à créer, et trop souvent dans des tableaux Excel…

Comment s'y prendre plus rationnellement ?
La solution consiste à créer un fichier de personnes contenant uniquement les informations personnelles. (Attention : la date ou année de naissance, et non l'âge, car l'âge n'est pas une donnée permanente…). On crée ensuite un fichier de lieux qui peut être, selon la finesse de la description, un ensemble de maisons ou un ensemble de villages, voire les deux. On crée par ailleurs un fichier d'interviews, un fichier de professions, etc. Une erreur inverse à éviter serait de créer plusieurs fichiers représentant le même type d'objets.
Par exemple, dans un travail récent sur les accoucheuses traditionnelles en Inde, nous avons construit une base de données d'interviews d'accoucheuses et de médecins. Nous avons vite réalisé qu'il fallait créer un seul fichier avec une case à cocher en fonction de la profession, plutôt que de séparer les deux populations. Les questionnaires ont donc été confondus, avec des parties spécifiques à chaque profession. C'est en étudiant la partie commune que nous pouvons maintenant proposer une étude comparative entre les pratiques " traditionnelles " et " officielles".
Un autre aspect de l'atomisation consiste à séparer les informations dans des champs distincts, par exemple nom, prénom, titre, pour les rassembler ensuite automatiquement par le biais de champs calculés qui mettent ces informations bout à bout.

Créer une base de données

Les données étant " atomisées " comme décrit précédemment, on comprend qu'il est nécessaire d'établir des liens pour que chaque personne, par exemple, soit attachée à une maison, chaque maison à un village, chaque interview à une personne et à un lieu, etc. Sans négliger le cas où une interview concerne plusieurs personnes, une personne habite plusieurs maisons ou a plusieurs professions, etc. Créer ces liens revient à construire la base de données proprement dite. Cela s'effectue en deux étapes :

- Indexer les informations

Pour appeler une information d'un fichier vers un autre il faut une référence unique à la fiche ciblée. S'il s'agit d'une personne, on peut se servir de son nom ou de la concaténation nom-prénom (créée automatiquement dans un champ calculé), le cas échéant de son numéro de sécurité sociale, etc. Toutes ces méthodes présentent tôt ou tard des problèmes en raison d'homonymies, d'informations manquantes ou d'erreurs de saisie. La seule solution qui marche à coup sûr est la suivante : créer dans chaque fichier un champ numérique qui inscrit automatiquement un numéro d'identification unique pour chaque nouvelle fiche créée. Empêcher la modification manuelle de cet identificateur.

- Créer les liens

Supposons que les villages ont ainsi été numérotés. Comment " attacher " un individu à un village ? On crée dans le fichier d'individus un champ numérique dans lequel on placera le numéro de village. On crée ensuite une relation, désignée par exemple comme " individu-lieu", qui met en contact deux fiches à condition que les numéros d'identification de lieux soient égaux. On peut ainsi créer de multiples liens entre de multiples fichiers. D'autre part, si l'on veut attacher le même individu à plusieurs villages, il suffit de créer dans le fichier d'individus un champ à valeurs multiples pour les numéros de chaque village. Les liens servent à afficher certaines informations provenant d'un autre fichier. Par exemple, dès que le numéro du lieu est entré dans une fiche, l'adresse complète peut s'afficher, ainsi que toute autre information contenue dans le fichier des lieux ou héritée, du district vers le taluka et du taluka vers le village.Cet héritage " vertical " est une propriété caractéristique des bases de données relationnelles.

L'approche orientée objet permet principalement de faire fonctionner d'autres formes, plus subtiles, d'héritage " latéral". Des relations multiples (one-to-many) peuvent être affichées. En créant, dans le fichier de lieux, un " portal", on peut faire apparaître par exemple la liste et les professions de toutes les personnes habitant dans chaque lieu. Dans une base de données relationnelle, l'information distante n'est pas copiée. Elle est simplement affichée. Si l'on veut effectuer des recherches sur cette information (par exemple, tous les individus qui habitent tel ou tel district) on fera appel à une variante : le modèle lookup dans lequel l'information est effectivement copiée (et peut être indexée pour une recherche rapide). Les inconvénients de cette solution sont que les informations copiées peuvent tenir de la place et que lorsqu'on modifie la source d'informations il faut lancer manuellement un " relookup " pour mettre à jour ses copies.

Des champs obligatoires

Dans tout fichier il est indispensable de créer deux champs date à entrée automatique, le premier indiquant la date de création de la fiche et le second celle de sa plus récente modification. Il sera facile ainsi de contrôler les saisies effectuées pendant une période donnée si l'on a des doutes sur certaines données. Si plusieurs personnes participent à la saisie des données, un champ servant à identifier l'opérateur permettra par la suite de le consulter au besoin.

Les tests de cohérence

Il est très important de détecter les erreurs de saisie dans une base de données. Cela peut se faire à trois

niveaux :

- le typage de données Pour entrer une date, utiliser un champ de type " date " plutôt qu'un champ texte ordinaire. Idem pour les champs numériques.
- la vérification de valeurs Si l'on entre une année de naissance, on vérifie par exemple que l'âge de la personne est dans une fourchette acceptable.
- la vérification des liens Il est indispensable de contrôler l'exactitude des numéros d'identification qui servent à attacher une fiche à des informations distantes.

Pour cela, par exemple, dans le fichier de personnes, on fera afficher le nom du village dès que son numéro aura été entré. Des vérifications plus sophistiquées peuvent être envisagées.
Dans la base de données sur les chanteuses du Maharastra sur laquelle j'ai travaillé au CCRSS de Pune, il y a un fichier PERFORMERS qui contient les interprètes, chacune liée à un village ou hameau du fichier LOCATIONS. Il y a d'autre part un fichier RECORDINGS qui liste les enregistrements, dans lequel est aussi indiqué le lieu de l'enregistrement. Enfin, le fichier SONGS contient les paroles de chaque chant, lié à la fois à une interprète et à un enregistrement sonore. Pour chaque chant, la base de données va chercher le lieu de l'interprète et le compare avec celui de l'enregistrement. Ils sont en principe identiques. Dans le cas contraire, un message est affiché. Ce test de cohérence nous a permis de détecter, et de considérer à part, les interprètes que nous avions enregistrées dans un lieu autre que leur village. Sans ce test nous étions exposés à des erreurs d'interprétation sur les particularités locales des chants.

Une base de données anthropologique

Cet exemple est issu d'une discussion informelle avec une doctorante en anthropologie qui avait bien du mal à démêler son écheveau de données. Je ne sais pas si elle a suivi ma proposition, mais je vous la livre à l'état brut. La base de données se compose de trois ensembles :

Des données de base : individus, lieux, castes, objets, toutes stockées dans des fichiers distincts. Il s'agit en principe des données " objectives " de l'enquête ;

Des récits, soit par des informateurs, soit par le chercheur (qui est aussi, à juste titre, un informateur), d'événements ayant eu lieu à des dates plus ou moins précises ;

Des points de vue sur les données de base et les récits, qui sont en fait des tentatives d'interprétation des données par les informateurs et le chercheur.

Chaque fiche de la base de données de récits peut contenir de nombreuses informations. Elle correspond à un témoignage collecté à une date précise. Elle contient donc le récit proprement dit, un résumé éventuel, un champ à valeurs multiples indiquant les auteurs du récit, un autre précisant les personnes ayant écouté ce récit, un champ pour le (s) lieu (x) de l'événement et un autre pour celui de la collection du récit, etc.
Ces liens vers les données de base permettent d'afficher automatiquement autant de détails sur le contexte de l'événement et celui du témoignage sur cet événement, deux notions qu'il est essentiel de bien dissocier dans une étude anthropologique. Le fichier de points de vue est encore plus riche en informations car il contient, outre le point de vue transcrit, son ou ses auteurs, des liens vers des récits, des lieux, des personnes, des objets…

C'est ici qu'il devient intéressant de travailler avec un logiciel permettant d'ajouter ou de modifier des champs aux moindres frais. Au fur et à mesure de la saisie, on s'aperçoit en effet qu'il est nécessaire de restructurer le fichier : si un point de vue est " lié " à une personne, par exemple, il faudra que la nature sémantique du lien soit précisée. C'est précisément ce travail de structuration de la base, devenu beaucoup plus souple avec les logiciels récents, qui constitue un fructueux pré-traitement des données et

permet en définitive de clarifier la méthodologie même de la recherche. Il est indispensable de l'accomplir sur le lieu même de la collection de données afin de pouvoir " corriger le tir " dès le commencement de l'enquête.

Ma réponse est donc : " oui", il est utile et parfois nécessaire d'emporter son ordinateur sur le terrain, dès lors qu'il sert à mieux qu'une machine à écrire… On peut le faire sans être informaticien si l'on dispose de quelques " tuyaux " techniques (voir en annexe).
L'informatique, c'est après tout une manière systématique de gérer de l'information (textuelle aussi bien que numérique), autrement dit un travail préliminaire indispensable à la recherche en sciences humaines et sociales…

_____ -_____.

Annexe : des précautions techniques

Sur les conditions techniques à réaliser pour éviter que l'enquête vire au scénario catastrophe, je me contenterai de rappeler quelques points qui me paraissent méconnus de nombreux jeunes utilisateurs de l'informatique :

Eviter de partir avec une machine flambant neuve. Un ordinateur a en principe besoin d'un an ou deux de fonctionnement pour révéler ses " défauts de jeunesse".

La machine n'a pas besoin d'être très rapide. Un modèle un peu ancien acheté d'occasion évite de grever son budget. Par contre il faut la munir d'assez de mémoire vive (32 ou 64 méga-octets ne sont pas un luxe), et le cas échéant changer le disque dur pour disposer d'une capacité suffisante (au moins 500 méga-octets). Installer un système " propre", récent, et toutes les précautions antivirus de dernière date.

Prévoir un dispositif de sauvegarde quotidienne. Les disquettes ont une durée de vie proche de zéro dans un environnement chaud, humide ou dans l'atmosphère polluée d'une ville comme Delhi. Il est donc indispensable d'utiliser un support de stockage fiable. Ils existent aujourd'hui à des prix abordables : Zip drive, Syquest, etc.

Les batteries rechargeables ont-elles aussi une durée de vie très courte dans des conditions extrêmes de température. Si l'ordinateur le permet, et en fonction du budget, on peut s'équiper de batteries externes au lithium, d'un panneau solaire ou tout simplement de batteries au plomb miniatures que l'on trouve chez tous les électriciens en Inde… Dans ce dernier cas il faut se munir d'appareils qui fonctionnent sous 6/12 Volts ou d'un adaptateur pour automobile.

Emporter les utilitaires de maintenance les plus importants : disquettes de démarrage avec un réparateur automatique de disques, un défragmenteur, ainsi qu'un diagnostic en tâche de fond dans le style de FileSaver des Symantec Norton Utilities.

Mettre dans son sac un tournevis spécial pour ouvrir la machine, ainsi qu'une souris en cas de blocage (fréquent) des boutons de clavier ainsi que, si vous optez pour un Mac, une paire de boîtiers de connexion AppleTalk…

Bernard Bel, le 23 mars 1998