III / De la source aux données

La transformation de la source « brute » en données quantifiables passe par deux grandes étapes : d'une part, la saisie, qui consiste à retranscrire l'archive au sein d'un document informatisé, d'autre part, le codage, qui modifie les informations recueillies lors de la saisie pour constituer des catégories, plus ou moins homogènes, afin de rendre possible un traitement quantifié. La confusion de ces deux étapes est à l'origine de nombre de critiques adressées à la quantification. Il convient, sans nul doute, de les distinguer clairement.

La saisie en pratique

Distinguer saisie et codage

Depuis l'introduction des micro-ordinateurs, le traitement statistique de données historiques en lui-même (obtention des résultats d'une analyse factorielle, d'une régression, etc.) peut en général se faire très rapidement. Même les traitements de gros corpus en lexicométrie et en analyse de réseaux, qui pouvaient prendre une journée vers 2000, se font le plus souvent en quelques minutes en 2007. Cela permet d'expérimenter différents traitements, de jouer sur les paramétrages et les recodages, donc d'affiner les résultats tout en sortant d'un positivisme du chiffre unique.

La mémoire et la rapidité des ordinateurs présentent deux autres avantages essentiels, qui permettent en partie de répondre aux critiques envers une catégorisation trop violente et

objectiviste, souvent encore vue comme une caractéristique obligée du quantitatif.

Premièrement, aucun logiciel ne limite plus la saisie de chaque « champ » à huit caractères ; en général, il est au moins possible d'en inclure quelques centaines, donc de respecter le langage et les détails de la source. Du coup, la saisie, comme étape du travail de recherche, doit impérativement être distinguée du codage. Si la première doit rester au plus proche de la source et viser une retranscription aussi fidèle que possible du document original, le second peut se faire plus tard, hors des lieux d'archives, et surtout se refaire autant de fois que nécessaire. De ce fait, il n'y a plus *un* mais *des* codages, plus ou moins fins selon ce que les techniques statistiques autorisent.

C'est le second avantage des tableurs, logiciels de bases de données et de statistiques, que de permettre très vite de recoder une variable, grâce soit à des fonctions dédiées, soit tout simplement aux outils de tri ou de recherche/remplacement. Il est tout aussi facile de changer les noms des modalités : le « 1 », « 2 », « 3 » non ambigu sied parfois mieux aux calculs, quand un graphique à publier se doit d'être plus explicite pour être lisible…

Mais toutes ces réflexions sur le codage optimal n'ont pas leur place lors de la saisie — sauf à être notées dans un coin comme bonnes idées à utiliser plus tard. Il convient au contraire d'adopter, face à la source, une approche large, pour éviter au maximum d'avoir à revenir en arrière. Certaines informations qui ne présentent pas d'intérêt à première vue peuvent devenir, au fil de la recherche, des indicateurs importants. Mieux vaut les saisir, sans coupe ni codage — ce qui ne dispense pas de réfléchir à la bonne manière de les saisir pour pouvoir ensuite les utiliser.

La saisie comme moment de recherche

Si codage et traitement peuvent donc se faire assez vite, de façon répétée et non continue, la saisie reste le véritable goulet d'étranglement. D'où la nécessité d'ajuster finement la taille de l'échantillon à la fois aux possibilités concrètes du travail sur la source (qui changent drastiquement selon qu'elle est déjà en ligne ou bien conservée dans des archives rarement ouvertes…) et à ce que l'on veut, *in fine*, démontrer. N'hésitez pas, par ailleurs, à multiplier les sauvegardes de vos documents de saisie, en les archivant aussi souvent que possible sur des supports

distincts de votre ordinateur (clé USB, CD…). La perte accidentelle de ces fichiers représente très souvent une catastrophe !

Promouvoir la saisie de la source par le chercheur ne correspond pas au désir d'imposer un « rite de passage » pénible et ingrat, mais bien à la conviction que cette étape constitue un véritable moment de recherche. La saisie n'impose plus de choix de codage irréversibles — du moins si l'historien ne se les impose pas à lui-même —, mais elle reste un moment important : fastidieuse, elle permet aussi d'appréhender réellement les données et de commencer à réfléchir sur leur structure. Le fait qu'elle ne puisse le plus souvent pas être sous-traitée, dans les conditions actuelles de la recherche, n'est donc pas seulement un inconvénient.

De plus, « l'ordinateur introduit une exigence de rigueur formelle en amont même du traitement » [Genet, 1986]. En effet, rester fidèle, le plus possible, au langage de la source n'empêche pas de devoir l'inscrire dans des lignes et dans des colonnes. L'opération n'a rien de trivial ; elle impose le respect de quelques principes facilitant les traitements ultérieurs, mais elle aide aussi à s'interroger sur la source et sur l'objet de recherche.

C'est souvent au cours de la saisie que l'on prend presque physiquement contact avec son sujet, mais aussi que naissent de nombreuses questions (qu'il est bon de noter, au fur et à mesure, sur un document à part). Ainsi, derrière l'aridité des documents administratifs destinés à l'identification et à l'enregistrement, se cachent des histoires individuelles dont le dépouillement permet de reconstituer les différentes étapes. Bien souvent présenté comme rébarbatif, ce temps de la saisie s'éclaire parfois de découvertes et de surprises. La saisie pour l'historien peut ainsi être comparée au terrain pour l'ethnographe ou le sociologue [Beaud et Weber, 1997] : elle engage physiquement, induit une connaissance intime de la source et suscite nombre des questionnements de recherche.

Répartir de l'information dans des lignes (par individu, qu'il s'agisse de personnes physiques, de livres ou d'entreprises) et des colonnes (par variable) impose de bien définir ces éléments. Il peut paraître très simple d'affecter une ligne à chaque individu, surtout lorsque l'on pense à des personnes physiques. Pourtant, on s'aperçoit parfois que l'unité à laquelle se rattachent nos informations est plutôt cette personne *dans une certaine source*,

Tableur ou gestionnaire de bases de données ?

Les traitements statistiques des données historiques se font en général soit dans un tableur (comptages, calculs de pourcentages, élaboration de graphiques) [Cellier et Cocaud, 2001 ; Saly-Giocanti, 2005], soit dans un logiciel dédié (de statistiques en général, ou bien d'analyse de réseaux, de données longitudinales, etc.). L'impératif catégorique, au moment de la saisie, consiste à produire des données transférables d'un logiciel à l'autre. Elles seront non seulement recodées, souvent simplifiées, mais aussi converties avant de conduire à des résultats, tandis que la base d'origine sera conservée à part, comme référence la plus proche de la source.

Fort heureusement, la plupart des logiciels reconnaissent les structures les plus simples : lignes et colonnes d'un tableau, la première ligne donnant les intitulés des colonnes, ou même *texte tabulé* (où un retour de chariot matérialise le changement de ligne, un espace ou une tabulation le changement de colonne). *A priori*, la saisie peut donc se faire à peu près n'importe où, même dans un traitement de texte.

Cependant, en pratique, le choix est souvent entre tableur et logiciel spécifique de bases de données. Ces derniers ont la préférence de beaucoup d'historiens du fait du caractère plus avenant de leurs formulaires de saisie — cependant, il est possible d'en obtenir du même type dans les tableurs. En fait, dans la plupart des cas, un tableur, plus souvent présent à la base sur l'ordinateur et d'un accès plus intuitif, suffit largement, d'autant que les données saisies ainsi sont plus facilement « exportables » par la suite, pour des traitements que l'on n'imagine que rarement à l'avance. Outre le gain de temps, la discipline imposée par le fait d'entrer les données directement dans un tableau, visualisé comme tel, peut d'ailleurs aider à poser certains choix de recherche.

Le bon usage des fonctions de largeur, d'affichage et de masquage des colonnes, d'aperçu avant impression, de tri et de filtre permet à la fois de gérer de grandes bases sans s'y perdre et d'utiliser le tableur pour des *requêtes* documentaires simples (afficher tous les individus possédant telle et telle caractéristique) autant que pour la saisie, le codage et les calculs. L'utilisation d'un tableur permet, en outre, un contrôle facile de la saisie : elle limite les risques d'erreur grâce à la détection rapide d'intitulés ou de valeurs aberrantes, pour les variables quantitatives par exemple.

Quant aux logiciels de bases de données, ils sont adaptés lorsque l'on dispose de données portant sur plusieurs entités de nature différente, mais que l'on souhaite étudier ensemble. Par exemple, on peut avoir des données à l'échelle du salarié et à l'échelle de l'entreprise, ou encore de l'auteur et de ses livres. Dans ce cas, il faut un tableau de données par type d'entité (un pour les salariés, un pour les entreprises) et une circulation souple entre l'un et l'autre. C'est possible, mais difficile, avec un tableur : il vaut mieux alors se former aux gestionnaires de bases de données, tout en sachant que, pour faire des calculs, il faudra exporter les données ailleurs [Cellier et Cocaud, 2001].

Un exemple de saisie dans un tableur

Iden-tifiant	N° registre	N° imma-triculation	Jour immat.	Mois immat.	An immat.	Nom commerce	Nom commer-çant	Nom j.f.	Prénom
1	573	504746	27	4	1931	Sastre	Sastre		Jean Eugène
2	573	505380	3	5	1931	Labbé	Labbé		Louis Gaston
3	574	506667	15	5	1931	Remacle	Remacle	Valleroy	Émilie Jeanne
4	574	506758	17	5	1931	Jarrige	Jarrige		Jean
5	574	507030	18	5	1931	Meviel	Meviel		Gabriel Léon
6	574	507550	25	5	1931	Buda	Buda		Jean
7	575	507705	26	5	1931	Charlon Legay	Charlon	Legay	Jeanne
8	575	508173	31	5	1931	Dunand	Dunand		Roger
9	575	508357	2	6	1931	Naïm	Naïm		Elias
10	576	510156	19	6	1931	Au bon Picolo	Hemon		Mathurin
11	576	510267	21	6	1931	Le point d'interro-gation	Girard		Charles
12	577	511313	1	7	1931	Blatanis	Blatanis		Constantin
13	577	511687	6	7	1931	Gohin	Gohin		André
14	578	512450	16	7	1931	Salon école de coiffure	Chanut		Gaston
15	578	512457	16	7	1931	Buvette de la Cité	Lepan	Paille	Louis Leopoldine
16	578	512762	20	7	1931	Grosjean	Grosjean		Nicolas Ernest
17	579	514066	3	8	1931	Malagnon	Malagnon		Albert
18	581	516831	6	9	1931	Mme Lanaud	Lanaud	Petiot	Françoise Claudine Berthe

Source : registre du commerce de la Seine, exploité dans Zalc [2002].

Sexe	Nationalité	Objet du commerce	N°	Rue	Arr/commune	Commune naiss.	Dpt naiss.	Jour	Mois naiss.	An. naiss.
M	Fr	transactions financières, comptabilité assurances	11	impasse de la Loi	20	Paris	Paris	11	12	1894
M	Fr	boucherie triperie	106	rue Vercingétorix	14	Montreuil-sous-Bois	Seine	29	4	1878
F	Fr	bonneterie layettes		forain		Plaine-St-Denis	Seine	1	3	1903
M	Fr	beurre œufs alimentation articles de ménage	28	rue Jean-Jaurès	Villejuif	St-Hilaire-Faissac	Corrèze	2	11	1891
M	Fr	restaurant buvette	7	rue de l'Étoile	17	St Léons	Aveyron	5	7	1898
M	Tchécoslovaque	peintre décorateur	12	rue du Pré-St-Gervais	19	Sudine	Tchécoslovaquie	19	7	1900
F	Fr	cycles électricité accessoires	40	av. de la Reine	Boulogne	Aix-Houlettes	Pas-de-Calais	29	9	1897
M	Suisse	charcuterie	121	rue de Paris	St Denis	Genève	Suisse	7	3	1901
M	Libanais	imprimerie	126	rue Pierre-Jaigneau	Bois Colombes	Chipoh	Liban	6	5	1903
M	Fr	vins restaurant	85	rue des Entrepreneurs	15	Juern	Morbihan	12	3	1881
M	Fr	café buffet froid		exposition coloniale		Durtal	Maine-et-Loire	29	1	1880
M	Héllène	fruits primeurs	158	rue Montmartre	2	Janina	Grèce	15	1	1899
M	Fr	chemises bonneterie divers	27	rue de Noisy	20	Paris	Paris	28	1	1900
M	Fr	salon, école de coiffure	20	rue Réaumur	3	Gueugnon	Saône et Loire	24	2	1879
F	Fr	épicerie buvette	90	rue Raymond-Poincaré	Nanterre	Nanterre		16	5	1899
M	Fr	maréchal-ferrant	141	av. de Paris	Genevilliers	Géricourt	Meuse	25	7	1881
M	Fr	vins liqueurs	20	rue Myrrha	18	Le Raincy	S et O	7	6	1899
F	Fr	chapeaux		ambulant		Bologny	Hte Marne	3	7	1874

Un exemple de source nécessitant l'usage d'un gestionnaire de bases de données

La thèse de Séverine Sofio sur les femmes peintres actives à Paris entre 1789 et 1848 [Sofio, 2007] se fonde en partie sur des catalogues d'exposition. Ceux-ci fournissent des données à la fois sur les artistes (statut matrimonial, adresse…) et sur ce qu'elles exposent (support, genre…). La plupart des artistes exposent plusieurs peintures ou autres objets d'art. Si on entre ces données dans un tableur avec une ligne par artiste, on doit « écraser » la diversité de ces œuvres, soit en les regroupant dans une seule case (« titres des œuvres »), soit en réalisant des comptages ou regroupements dès la saisie (« genre majoritairement pratiqué »). Si on utilise un tableur, mais avec une ligne par œuvre, on doit répéter de façon fastidieuse les informations sur chaque artiste, et surtout on ne peut pas compter facilement le nombre d'artistes (global ou selon le statut matrimonial, l'adresse, etc.).

Dans ce genre de cas, les gestionnaires de bases de données présentent un réel intérêt. Ils permettent en effet de créer une *table* (qui correspond à une feuille de saisie dans un tableur) pour les artistes, une autre pour les œuvres, et de *lier* ces deux tables par le biais d'identifiants numériques, pour que l'on sache à quel artiste de la première table correspond chaque œuvre de la seconde. On peut ainsi réaliser commodément tant des comptages ou tris séparés portant sur les artistes ou sur les œuvres que des requêtes complexes interrogeant les deux tables à la fois (comme « combien d'artistes de la rive gauche peignent des fleurs ? »).

dans un certain rôle ou *une certaine année*, et qu'il est plus intéressant de présenter ainsi les données, au moins dans un premier temps, avec plusieurs lignes pour chaque personne physique.

Dans ce cas, bien sûr, une colonne devra indiquer à quelle personne physique (désignée par son nom et de préférence par un numéro) se rapporte chaque ligne, pour autoriser un regroupement ultérieur. Même pour les sources *a priori* les plus simples, ressemblant déjà à un tableau, comme une liste de membres d'une association, on s'aperçoit souvent lors de la saisie que chaque ligne ne décrit pas forcément une personne physique, mais parfois une entreprise ou une autre association : selon l'objectif de la recherche, il faut alors s'interroger sur le meilleur mode de saisie.

Quant aux colonnes, si quelques impératifs pratiques peuvent guider leur découpage, elles renvoient aussi à des questions de fond sur la définition de l'information fournie par la source. Par exemple, si une catégorie « qualité » de la source regroupe, de notre point de vue, des informations hétérogènes (professions,

Les dix commandements de la saisie

Remarque liminaire : la plupart de ces recommandations visent à segmenter au maximum l'information lors de la saisie. Il est en effet techniquement plus simple, ensuite, de regrouper (de façon en général automatique) que de scinder (de façon en général manuelle) le contenu des cases. Éclater les informations facilite les opérations de tri (qui se font sur le premier caractère de la case). Cela conduit à multiplier les colonnes, ce qui ne doit pas, en soi, poser de problème. Si l'on dépasse les limites en la matière d'un tableur, il est toujours possible d'ouvrir une nouvelle « feuille de calcul », tant qu'elle renvoie bien aux mêmes individus.

1. Réserver une et une seule ligne, la première, aux intitulés des variables.

2. Utiliser une colonne pour attribuer des identifiants numériques arbitraires aux individus étudiés (individu n° 1, n° 2…). Ils servent à éviter les ambiguïtés liées à la typographie (accents…) ou à l'homonymie et parfois à relier des informations issues de différentes bases. Il n'est pas gênant que certains chiffres ne soient pas ou plus attribués comme identifiants (suite par exemple à la découverte que deux individus ne font en fait qu'un).

3. Conserver le plus possible les formulations de la source. S'il faut simplifier, moderniser l'orthographe, etc., le faire *a posteriori*, dans une autre colonne ou un autre fichier.

4. Conserver le lien avec la source. Selon qu'une source est à l'origine de toute la base, de toutes les données pour une personne, pour une variable, d'une seule donnée, etc., les solutions varient, mais la référence de la source doit apparaître. On peut ainsi avoir une source par *classeur*, une source par *feuille* d'un *classeur*, une source par ligne ou colonne (indiquée au début de celle-ci), une colonne « source des renseignements de la colonne précédente », etc. L'essentiel est de conserver l'information précise

mandats et décorations), il peut être utile, lors de la saisie, de garder le lien avec la source tout en facilitant les traitements analytiques ultérieurs. On créera alors trois colonnes (qualité-profession, qualité-mandat, qualité-décoration) ou quatre (il y aura sans doute aussi « qualité-autre ») plutôt qu'une seule.

Catégories et codages

Le codage est l'un des points sur lesquels s'est focalisée la critique des méthodes quantitatives depuis les années 1980 : transposition de nomenclatures anachroniques, simplification des données, réification des individus, agrégation du divers, il se

dans au moins une version de la base de données, quitte ensuite à procéder à des regroupements et simplifications.

5. Utiliser les fonctions d'annotation ou de commentaire des logiciels, par exemple pour mentionner la source particulière d'un renseignement, un choix de transcription ou de codage, un doute sur une graphie…

6. Scinder le plus possible l'information, par exemple avec des colonnes « M./Mme/Mlle », « nom », « prénom 1 », « prénom 2 », « prénom 3 », « titre de noblesse », « nom de jeune fille » et « pseudonyme » plutôt qu'une seule colonne « nom ». De même pour les adresses : non que le numéro soit assez intéressant pour constituer une variable, mais parce que le mettre dans une colonne à part permet de trier par nom de rue.

7. Conserver le lien avec la date de l'information. Si la profession est connue à plusieurs dates, faire une colonne par date, ou des colonnes « profession 1 », « date de la profession 1 », « profession 2 », « date de la profession 2 », etc. Si les dates sont en fait des intervalles, faire une colonne pour la date de début et une pour la date de fin.

8. Éviter, sauf si on travaille seulement sur le XX^e^ siècle, le format « date » des logiciels, difficilement exportable. Préférer le codage en trois colonnes : jour, mois et année (en format « nombre »), qui permet des tris plus fins.

9. Aller explorer les fonctions suivantes dans l'aide du tableur : figer les volets, ajustement automatique de la largeur des colonnes, sélections multiples, copie automatique ou incrémentation, collage spécial, remplacer, insertion de fonction, concaténer, mise en pages, zone d'impression, tri, filtre, tableau croisé dynamique ou pilote de données ; et regarder tout ce que l'on peut faire avec un clic droit sur un PC (ou Ctrl + clic sur Macintosh).

10. Sauvegarder aussi souvent que possible.

heurte en pratique à de nombreux écueils. De plus, il mutile, pour une part, la richesse de l'information extraite du dépouillement brut de la source ; mais il apparaît nécessaire afin de réaliser un traitement statistique des données. Surtout, on peut également concevoir cette étape comme un moment de réflexion sur les sources et sur l'objet, qui permet d'expliciter les problèmes de dénomination et de comparabilité.

Revenir aux catégories indigènes ?

La déconstruction des catégories statistiques, en histoire, part d'une critique de l'importation dans le passé de constructions toutes récentes, comme les agrégats de la comptabilité nationale, les catégories socioprofessionnelles de l'Insee ou encore le concept de chômage [Desrosières et Thévenot, 1988 ; Charle, 1993 ; Topalov, 1994]. La critique exagère d'ailleurs parfois les

tentations en la matière des historiens des années 1960 et 1970, qui se posaient souvent de vraies questions de codage et ne choisissaient pas toujours l'anachronisme [Garden, 1970].

Que faire après cette déconstruction ? Elle a donné naissance à une histoire de la statistique, préoccupée de l'évolution intellectuelle, matérielle et politique de ses méthodes et de ses institutions [Desrosières, 1993]. Néanmoins, ce courant bien vivace s'est, en général, nettement éloigné de l'histoire quantitative. Pour ceux qui souhaitent tout de même compter, donc classer, les choix sont plus difficiles. Un mot d'ordre répandu est le « retour aux catégories indigènes » : celles, non anachroniques, de l'époque, des sources, des acteurs.

Mais il a rapidement fait lui-même l'objet de critiques. Celles-ci sont par exemple formulées par Daniel Milo [1987] à propos d'études sur les pratiques de lecture du XVIII[e] siècle. Une série de travaux en la matière, fondés sur des inventaires de bibliothèques, ont reproduit un classement thématique en cinq catégories qui était présenté par François Furet comme « établi selon les critères de l'époque ». Cependant, d'autres études quantitatives ont utilisé des catégories différentes, souvent plus fines. En réalité, les sources étaient loin d'imposer ou même de proposer des critères clairs de classement.

En effet, aucune époque, aucun groupe n'ont connu de « catégories indigènes » consensuelles et évidentes. Adopter un classement plutôt qu'un autre peut ainsi conduire à occulter certains conflits, voire à masquer certains groupes. C'est plutôt la confrontation de différentes catégories indigènes qui permet d'une part d'approfondir l'étude qualitative des enjeux des classifications, d'autre part de faire à bon escient des choix de codage, s'ils sont requis par une visée de quantification.

Le codage, un choix souvent politique

Il est ainsi important de distinguer les identités sociales déclarées par les individus, dans telle ou telle situation, des propriétés que l'historien leur attribue, de façon parfois très décontextualisée [Cerutti, 1995]. Le mot même d'« identité », parfois piégé en raison de ses connotations fixistes et objet de confusions fréquentes entre identification par une autorité, appartenances revendiquées par les individus et image sociale, est mis en débat depuis les années 1990 [Avanza et Laferté, 2005].

Statistiques « ethniques » et autocatégorisations

La question très controversée des statistiques « ethniques », dont certains demandent la mise en place en France, depuis la fin des années 1990, pour lutter contre les discriminations, comporte de multiples enjeux [Merllié et Spire, 1999]. Les historiens de la statistique ont montré la variété des méthodes employées, dans le passé, pour « compter l'autre » et leur signification politique [*Histoire & Mesure*, 1998].

Un des apports de ces nombreux débats et études est de remettre en cause la solution naïve aux questions de catégorisation qui consiste à laisser les individus se classer eux-mêmes. Les manières d'énoncer son appartenance ne sont pas toujours formatées par les catégories juridiques. Elles dépendent également des situations, lieux, périodes de recueil de l'information. Face à la question de leur « nationalité », certains commerçants du département de la Seine dans l'entre-deux-guerres donnent des réponses étonnantes (« géorgien grec », « brésilien arménien »). Les mentions « israélite » ou « israélite du Levant », retrouvées à plusieurs reprises, sont ainsi significatives d'une appartenance nationale vécue sur un mode religieux, propre aux juifs orientaux originaires de l'Empire ottoman où, jusqu'en 1839, l'État conférait à chaque communauté religieuse non musulmane une autonomie juridique. D'autres indiquent une nationalité « indéterminée » ou invoquent le statut qu'ils revendiquent (« réfugié russe ») [Zalc, 1998].

Tout codage implique un choix, donc réduit les nuances de la source tout en imposant des modèles d'analyse. La distinction fréquente, reflétant bien souvent les données disponibles concernant l'immigration, entre « Français », « naturalisés » et « étrangers », engage un implicite : ne pas considérer les « naturalisés » comme des Français. Le choix des termes pour qualifier telle ou telle catégorie n'est pas neutre : « naturalisé » n'équivaut pas à « Français par acquisition » par exemple.

Face à ces difficultés, mais aussi à la nécessité d'un minimum de classement pour la présentation de résultats, comment dégager des principes opératoires ? En se proposant de faire une « histoire sociale du codage statistique », Alain Desrosières [1989] a souligné que les catégories se devaient avant tout de produire, en donnant forme au chaos des données, des « choses qui tiennent ». Les considérer comme des conventions ne doit pas, selon lui, mener au relativisme. Toute catégorisation prétend construire des classes d'objets équivalents d'un certain point de vue : un point de vue orienté vers une certaine forme d'action sur ces choses, elle-même liée à une certaine vision des causalités.

Laisser les données se classer elles-mêmes ?

La critique des catégories imposées de l'extérieur (du présent vers le passé, en particulier) en termes de violence faite aux données a fait naître l'utopie contraire, celle de données qui indiqueraient d'elles-mêmes au chercheur la meilleure façon de les classer.

Les techniques de « classification automatique » (*clustering*) en sont la traduction la plus pure. Elles visent à constituer des groupes à partir des données observées elles-mêmes, le moins précodées possible (même si les choix de corpus et de saisie pèsent évidemment sur le résultat). Plutôt que d'utiliser des critères extérieurs, explicités *a priori*, le chercheur laisse le logiciel déterminer les proximités ou distances entre individus à partir de l'ensemble des observations, puis créer des groupes de façon que les individus d'un groupe soient proches entre eux et éloignés de ceux des autres groupes. Ces techniques peuvent produire des indices pertinents de regroupement et de différenciation. Mais elles sont encore peu utilisées, sauf dans certains domaines où elles font figure d'auxiliaires (lexicométrie, analyse de séquences : *cf.* chapitre VI).

En effet, en pratique, les choix de paramétrage sont nombreux et peuvent conduire à des résultats variés, entre lesquels il est difficile de trancher. La communauté, encore restreinte, des chercheurs qui utilisent ces méthodes commence seulement à se mettre d'accord sur de bonnes pratiques propres à assurer une plus grande stabilité et une meilleure lecture des résultats. Car, même pour une machine, il n'y a pas de catégorisation parfaite dans l'absolu…

Ce qui vaut pour l'action politique vaut aussi pour la recherche : il n'y a pas de catégorisation parfaite en elle-même, mais il en est qui sont plus ou moins adaptées à des objectifs de recherche, que ce soit du point de vue pratique (nombre de classes, par exemple), théorique (critères de classement) ou rhétorique (noms donnés aux groupes obtenus). « La question n'est pas : "Ces objets sont-ils *vraiment* équivalents ?" mais : "Qui décide de les traiter comme équivalents et dans quel but ?" », souligne A. Desrosières. Il y a bien des façons de catégoriser : à partir de critères posés *a priori* certes, mais aussi, par exemple, en partant de cas types et en les rapprochant des autres.

Coder et recoder

Une fois la saisie réalisée et sauvegardée à part, le codage doit surtout se faire en fonction des modes d'exploration des données que l'on envisage. Toutefois, il constitue, en lui-même, une étape déterminante de la recherche, notamment quand il est particulièrement délicat, comme pour des classements

thématiques de titres d'ouvrages ou pour des regroupements de mentions très hétérogènes, de professions par exemple. Premier traitement de la source, l'exercice de codage permet alors de poser un certain nombre d'hypothèses, voire d'élaborer de premiers résultats.

Dans ces cas, il est utile de noter au fil de la saisie, dans une colonne à part, les premières idées de codage qui peuvent surgir, pas forcément d'ailleurs pour tous les individus. Un tri ultérieur sur cette colonne permet déjà de voir à combien de catégories on a pensé, si certaines sont très peu employées, très proches l'une de l'autre, etc. De façon complémentaire, on peut utiliser les fonctions de « tableau croisé dynamique » ou « pilote de données » du tableur sur une seule colonne pour obtenir la liste de toutes les mentions différentes qui y sont présentes et des effectifs pour celles qui se répètent (cela revient en fait à faire un *tri à plat*), ce qui permet de juger de l'hétérogénéité des données. Dans tous les cas, il ne faut pas hésiter à faire plusieurs tentatives successives de codage d'une même variable et à les conserver dans plusieurs colonnes : certaines peuvent être plus ou moins adaptées à tel ou tel traitement statistique.

Quant aux méthodes de codage elles-mêmes, il n'y a guère de principes généraux : mieux vaut au contraire se permettre d'être inventif. En particulier, tout codage n'est pas forcément une *partition* (division d'un ensemble en sous-ensembles étanches, de façon que chaque chose soit dans une et une seule boˆte). Il peut être intéressant, pour certaines variables, d'adopter un codage binaire autorisant plusieurs modalités pour la même personne. Par exemple, les membres de la chambre de commerce évoqués au chapitre II peuvent déclarer plusieurs activités. Une façon de coder consiste à créer des colonnes « banquier », « négociant », « marchand de vins », « libraire », etc., puis à les remplir, pour chaque individu-ligne, de manière binaire (oui/non). Cela permet ensuite de s'intéresser à chaque mention séparément, mais aussi d'utiliser des critères de regroupement originaux, opposant par exemple « une seule activité déclarée » à « plusieurs activités déclarées ».

À chaque source, mais aussi à chaque questionnement correspondent des codages. Ainsi, pour retracer les trajectoires d'immigrés arrivés dans le Cher durant l'entre-deux-guerres, P. Rygiel [2001] construit une classification socioprofessionnelle complexe élaborée en fonction de ses données et de ses

Deux exemples de codage

L'exploitation du registre du commerce [Zalc, 2002] montre bien les problèmes posés, mais aussi les apports de la saisie et du codage. À première vue, cette source fournit presque directement des renseignements « classiques », faciles à placer dans des cases, comme la nationalité, l'adresse ou l'objet du commerce. Cependant, au fil du codage, l'arbitrage est permanent entre perte de sens et gain d'intelligibilité.

Le codage de l'« objet du commerce » pose ainsi le problème du choix d'un degré de finesse. La diversité des types d'entreprises regroupées dans un même document, où se côtoient marchands d'« articles de Paris », courtiers en assurances, blanchisseurs, fabricants de ceintures, négociants en gros de fruits et légumes, etc. comme les modalités de l'enregistrement des activités placent l'historien face à une profusion de déclarations dont la précision est extrêmement variable. Certaines mentions relevées, comme « tissus », ne permettent pas de distinguer fabrication et commercialisation, alors que la différenciation artisanat-commerce faisait partie des questions de recherche initiales. Le codage s'est efforcé de rendre compte de la spécificité de la source en ne plaquant pas une grille de lecture façonnée *a priori* sur les objets du commerce. Le choix de certaines catégories, comme le fait de placer la coiffure à part, a en outre été dicté par une connaissance plus générale du contexte, faisant subodorer des comportements particuliers, plutôt que par des impératifs liés à des critères généraux ou aux effectifs.

De plus, le codage d'un élément inattendu, le nom de l'entreprise, a apporté des résultats importants. Il fallait ici inventer des catégories *ad hoc* : nom fondé ou non sur le patronyme de l'entrepreneur, renvoyant à un pays, un objet ou au lieu du commerce… Ce codage est orienté par une question de recherche précise : il était reproché aux étrangers, à l'époque, de tromper les clients en francisant leur nom de commerce. Le codage et le comptage invalident ce préjugé, tout en permettant une réflexion sur les stratégies publicitaires et les modes de gestion de l'origine selon les secteurs et les périodes.

problématiques. Cette construction d'un dispositif d'enquête adapté lui permet de formuler des hypothèses, mais également des résultats quant aux mobilités sociales différenciées des immigrés et de leurs enfants.

L'important est alors de rendre explicites pour les lecteurs les choix réalisés, de ne pas appliquer de nomenclatures préétablies, de multiplier les essais et les grilles de codage possibles.