LE CODAGE AUTOMATIQUE D'UN CARNET DE DEPENSES EST-IL PLUS COMPLEXE QUE CELUI D'UN CARNET D'ACTIVITES?

F. DESCHAMPS[*], *S. DESTANDAU*[**] *et F. DUMONTIER*[**]

[*] INSEE, Unité Méthodes Statistiques
[**] INSEE, Division conditions de vie des ménages

Introduction :

Dans cet article, nous vous présentons les préparatifs à la codification automatique avec Sicore[1] *de deux enquêtes auprès des ménages : Emploi du temps (EdT*[2]*) et Budget de Famille (BdF*[3]*) et leurs résultats.*

Ces deux enquêtes peuvent être considérées comme similaires pour plusieurs raisons : ce sont chacune des enquêtes qui cherchent à mesurer un phénomène saisonnier : l'activité et la dépense ; par conséquent, elles se déroulent sur 8 vagues de 6 semaines, au cours d'une année entière. De plus, ces deux variables sont renseignées en clair par les ménages interrogés. Ces deux enquêtes ont donc eu recours à la codification automatique par Sicore lors de leurs dernières éditions de 1998-1999 pour EdT et de 2000-2001 pour BdF[4] *: l'initialisation des bases de connaissances utiles à Sicore a dû être menée pour les deux variables concernées.*

Nous vous montrons qu'en dépit de fortes similitudes sur le dispositif des deux enquêtes, les travaux préparatoires et les bilans de codification (finaux pour EdT et provisoires pour BdF) sont bien différents du fait du type de variable à coder, de la nomenclature associée, et du dispositif de recueil de l'information.

[1] Pour plus de renseignements sur Sicore et ses connaissances, on se reportera à l'annexe n°1

[2] Pour plus de renseignements sur l'enquête Emploi du Temps, on se reportera à l'annexe n°2

[3] Pour plus de renseignements sur l'enquête Budget de Famille, on se reportera à l'annexe n°5

[4] L'enquête BdF est encore sur le terrain à la date de ces journées.

I. SICORE et l'activité journalière des personnes dans l'enquête Emploi du Temps 1998-1999.

1. La variable activité et son support de collecte : le carnet

1.1 Le carnet d'emploi du temps journalier

Le carnet journalier de l'enquête Emploi du Temps permet de décrire toutes les activités d'un même jour de certains individus du ménage interrogé. Celui de l'enquête de 1986 était découpé en 288 lignes représentant les tranches de 5 mn d'une journée (de 0h à 24h) ; un individu kish et son conjoint éventuel devaient remplir un carnet chacun.

Pour se conformer aux normes des enquêtes EdT européennes, il a été décidé que le carnet de la version de 1998-1999 serait découpé en 144 tranches de 10 mn. De plus, tous les individus du ménage de plus de 15 ans rempliraient un carnet de même type. Sur une ligne donnée correspondant à une tranche dans les 2 éditions (1986 et 1998-1999), l'individu devait, par un trait vertical, joindre l'heure/minute du début de son activité à celle de sa fin sur le carnet. Ensuite, l'activité était indiquée " en clair " sur une ligne à l'intérieur de la plage horaire définie précédemment.

1.2 Une variable homogène et structurée

L'enquêteur insiste auprès des ménages pour qu'ils décrivent leurs activités **de façon claire, par une phrase simple, dans leur langage courant** (exemple : je conduis les enfants à l'école) et sur le fait qu'il fallait que l'individu ne décrive qu'une activité à la fois. Nous avons obtenu des phrases structurées du type " sujet + verbe + (compléments) ".

Or, très souvent, plusieurs activités pouvaient être réalisées simultanément. On demandait alors aux enquêtés de n'en noter que deux. La plus importante des deux était alors appelée " activité principale " et inscrite sur une ligne du carnet avec sa durée. La deuxième dite " activité secondaire " était aussi inscrite sur la même ligne du carnet que la première, mais dans une autre colonne - moins large. La durée, par construction du carnet, correspondait à celle de l'activité principale associée.
Chaque plage horaire de 10mn tient sur deux pages face à face du carnet (cf. annexe n°3). La page de droite est réservée aux intitulés des activités primaires et secondaires, la page gauche, aux variables annexes. Sur la page de droite, on a délibérément offert deux fois plus de place pour les intitulés des activités primaires que pour les secondaires, de façon à laisser aux individus assez de place pour décrire correctement les premières. En revanche, on acceptait pour les activités secondaires une description plus sommaire (comme « radio », « conversation »).

Extrait de la page gauche d'un carnet de l'enquête EdT 1998-99

	Marquez vos différentes occupations de la journée en indiquant les heures de début et de fin de chaque occupation à l'aide d'accolades dans la colonne de gauche		**Faites-vous autre chose en même temps ?** *(lecture, conversation, radio, TV...)*
7h00			
10			
20		**je dors**	
30			
40		**je fais ma toilette**	**radio**
50		**je m'habille**	**radio**
8h00		**je prépare le petit déjeuner**	**conversation**
10		**je déjeune**	**TV**
20			
30		**je fais la vaisselle**	**TV**
40		**je range la cuisine**	**TV**
50		**je fais le ménage**	
9h00		**je vais au travail avec ma**	
10		**femme et un voisin**	**conversation**
20			

Cette présentation était la même lors des enquêtes précédentes.

2. La nomenclature initiale de la variable activité

La nomenclature de 1986 contenait 199 postes de base à trois chiffres, qui pouvaient être regroupés en huit grands postes correspondants au premier chiffre :

Premier chiffre	Intitulé du poste	Exemples
1	Besoins physiologiques	Sommeil, soins personnels (se laver, s'habiller..) soins médicaux, repas
2	Temps de travail professionnel et temps de formation	Travail professionnel, formation professionnelle, études (étudiants, lycéens), autres formations
3	Travaux domestiques	Cuisine, ménage, soins du linge, courses, services administratifs, bricolage, jardinage, divers...
4	Soins aux personnes	S'occuper des enfants, jouer avec, soins matériels ou médicaux aux adultes
5	Sociabilité	Réceptions et sorties, conversation, téléphone, courrier, religion, participation civique et entraide
6	Loisirs	Sport, promenade, chasse - pêche, médias, spectacles, passe-temps et jeux
8	Trajets	Trajets domicile-travail, trajets liés aux enfants, autres trajets
9	Remplissage du carnet Insee	

3. Pourquoi utiliser Sicore[5] pour coder l'activité?

Pour l'enquête précédente de 1986, le codage de l'activité figurant dans les carnets journaliers était effectué par une seule équipe de codeurs du CNE de Toulouse et pesait très lourd en heures manuelles de codage (45 000 heures environ).

Le CNE de Toulouse n'existant plus, et les DR ne disposant pas d'une telle ressource de moyens manuels pour que l'enquête puisse avoir lieu sous la même forme qu'en 1986, il fallait trouver un moyen de réduire ces charges. Nous avons alors étudié la possibilité d'utiliser Sicore. Cela constituait une première : l'activité quotidienne était une nouvelle variable jamais codée et, de plus, les activités des anciennes enquêtes n'avaient jamais été saisies sur support informatique.

Toutefois, une chose était sûre, nous aurions moins de cas à traiter manuellement que lors de l'enquête précédente. Mais nous ne savions pas, au tout début des préparatifs relatifs à l'utilisation de Sicore, si le codage avec Sicore serait d'aussi bonne qualité qu'un codage manuel qui aurait été effectué de la même manière qu'en 1986.

4. La constitution des connaissances

4.1 Le premier fichier d'Apprentissage Brut (Fab) des activités

Dans la documentation de l'enquête de 1986, pour donner un aperçu de la diversité des libellés et de la richesse de cette nomenclature, figurait une liste de 3 918 de libellés différents, issue de 900 carnets de l'enquête de 1986, chaque libellé étant associé à un poste de la nomenclature. Après la saisie de cette liste de libellés et de codes, un premier fichier d'apprentissage était né. Il comportait 3 918 lignes.

Afin de tester ce Fab, la technique utilisée par *Pascal Rivière*[6] a été :

- d'extraire un échantillon du Fab
- d'effectuer l'apprentissage sur le reste du Fab
- de coder cet échantillon.

[5] Pour plus de renseignements sur Sicore et ses connaissances, on se reportera à l'annexe n°1

[6] A l'époque, en mars 1995, Pascal Rivière était C.P.S. du projet Sicore, alors en phase d'achèvement. A cette occasion, il a formé Jean-Louis Pan Ké Shon, expert de la variable activité.

Dix-neuf échantillons ont été ainsi extraits et codés de cette façon. En moyenne, le taux moyen de réussite de ces 19 codages a été de 40.4% (avec un minimum de 35% et un maximum de 46%) dont 32,4% codés simples et 8% codés multiples.

La qualité moyenne brute, pourcentage de "bien codés" parmi les libellés codés automatiquement a été de 50,5%. L'expression de "bien codés" a ici un sens précis, dans la mesure où chaque libellé de l'échantillon est issu du Fab et par conséquent accompagné d'un code, que l'on appellera code de référence. Un libellé "bien codé" est un libellé dont le contenu obtenu automatiquement est égal au code de référence.

Mais, le fait qu'il existe des codes multiples pour un même libellé va conduire de açon mécanique à coder des libellés dans un code faux (i.e. différent du fichier d'apprentissage) et conduire à une très nette et artificielle diminution de la qualité du chiffrement automatique. En étudiant les codés multiples, *Pascal Rivière* a évalué à 10,9%, cette proportion de mal codés mécaniquement. Pour calculer la véritable qualité, il fallait tenir compte de cet effet pour redresser la qualité moyenne brute. Il a donc évalué une qualité redressée de 76%.

4.2 L'amélioration standard des connaissances pour la variable activité

La mise en œuvre de la boucle Sicore conduit à ce que l'expert variable[7] effectue toujours le même genre d'opérations sur les connaissances relatives à la variable :

- L'ajout d'expressions originales au sein du Fab, issues de la saisie de carnets soit de l'enquête EdT précédente de 1986 soit des tests ayant eu lieu sur le terrain[8].

- La déclinaison des libellés déjà présent dans le Fab.
Les déclinaisons verbales au sens le plus strict en font partie.
Par exemple, ayant " JARDINAGE ", on peut rajouter : " J'AI JARDINE ", " NOUS JARDINONS ", " JARDINANT ", ...
On peut également imaginer des variantes plus larges des libellés.
Par exemple, " REVEIL " peut amener à " MAMAN ME REVEILLE ".

- Le choix des caractères blancs ou vides et des mots vides.
Par exemple, supposons que toutes les déclinaisons du verbe être et tous les articles et prépositions sont éliminés à la normalisation (ce sont des mots vides). Si, de plus " MAISON " est un autre mot vide car il indique un lieu plus qu'une action, alors le libellé " JE SUIS A LA MAISON " devient un

[7] En l'occurrence *Jean-Louis Pan Ké Shon*

[8] Il y a eu trois tests terrain : un test méthodologique, uniquement sur les carnets, effectué à Strasbourg et à Dijon du 15 juin au 14 juillet 96 ainsi que deux tests grandeur nature de l'ensemble de l'enquête en juin 1997 et en septembre 1997.

libellé blanc, alors qu'il pourrait être codé en repos, ou en réflexion. Les mots vides doivent alors être revus.

- Le choix, l'intégration, la mise en ordre et l'élimination de synonymes inutiles. Le choix des synonymes se fait grâce à l'étude des fréquences des libellés soit mal codés par Sicore, soit non codés. Leur intégration dépend de la nomenclature et doit respecter un ordre à tester pour une utilisation optimale.

Par exemple, supposons que dans la liste des synonymes, figurent dans l'ordre toutes les déclinaisons de " DORS ", puis toutes celles de " SOMNOLE " ; ensuite, dans la liste des synonymes « SOMNOLE » sera associé à " DORS ", alors, quelle que soit la forme verbale du verbe somnoler, celle-ci sera codée comme " DORS ". Certains synonymes peuvent avoir un effet non envisagé initialement qui aboutit à leur suppression.

Ainsi, voici quelques caractéristiques respectives du premier et du dernier environnement de codage de la variable activité :

Environnement	Premier	Dernier[9]
Nombre de lignes du Fab	3 918	13 400
Nombre de synonymes	0	2 388

4.3 La mesure de la qualité du codage automatique

Etant donnés les petits volumes de libellés lors des 3 tests de cette enquête, il a été possible de demander un codage manuel aux personnes effectuant la saisie des libellés d'activité. Cela a permis d'avoir un code de référence que Sicore pouvait comparer au code qu'il attribuait, et donc d'évaluer une certaine qualité de codage automatique. De plus, l'expert variable a analysé les rejets et les codés multiples de la codification automatique pour améliorer l'environnement Sicore ainsi que le prévoit le dispositif de la boucle Sicore

[9] L'environnement qui a servi au recodage complet des libellés de l'enquête à la fin de celle-ci.

Les résultats de ces tests sont les suivants :

Date	Nombre de libellés	Envirronement SICORE	Codés Sicore	Qualité[10]
test méthodologique juin – juillet 1996	3 253	Fab = 9500 lignes 1000 synonymes introduction des mots joker	80%	90%
1er test terrain juin 1997	4 270	Fab = 10 500 lignes 1 650 synonymes	83%	93,5%
2ème test terrain[11] septembre 1997	4 220	1er codage même environnement que test précédent	78,5%	80,5%
2ème test terrain	4 220	2 ème codage environnement amélioré	81,6%	96,1%

Juste après le deuxième test terrain, la codification automatique par Sicore donnait plus de 80% de libellés codés avec une excellente qualité.

4.4 L'utilisation de variables annexes

Le codage du premier Fab sur lui-même a permis de détecter des libellés qui menaient à plusieurs codes. (1/5 des libellés - cf. annexe 4) .
Ceci a confirmé, ce qui était prévisible, au vu de la structure de la nomenclature, qu'il fallait introduire des variables annexes dans le carnet qui donnent des règles de codage pour Sicore.

> *Par exemple, " BOIS CAFE " pouvait être codé 146 (prendre le café ou le thé à domicile), 156 (prendre le café ou le thé sur son lieu de son travail), 163 (prendre le café, le thé chez des amis, voisins, parents lors de visites, réceptions) ou 166 (prendre le café, le thé à domicile, avec des amis, voisins, lors de réceptions, visites).*

En 1986, certaines variables annexes étaient relevées en clair sur le carnet dans deux colonnes séparées en face de chaque activité : " en présence de qui ? " et le " lieu de l'occupation ". Ces deux variables annexes se sont donc imposées d'elles-mêmes. Le nombre de modalités a été fixé à 5 pour la variable lieu (" chez soi ", " lieu de travail ", " à l'extérieur ", " trajet autre domicile-travail " et " autre trajet ") et à 4

[10] La qualité calculée ici est la proportion de codes parmi les codés automatiques, qui sont égaux aux codes manuels que les DR avaient codés manuellement avant la saisie.

[11] En septembre 1997, nous avons fait deux codages, l'un en gardant, sans l'enrichir, l'environnement de juin 1997, et l'autre en l'enrichissant au vu des résultats du codage de juin 1997. Nous avions alors obtenu ce que nous espérions, c'est à dire un taux final de codage dépassant 80% avec une excellente qualité (96%).

pour la variable présence (" seule ", " autre personne du ménage ", " amis, voisins, parenté, collègue " et " autre personne ").
En plus de ces variables, d'autres variables annexes ont pu être définies sur les carnets de 1986, et reprises en 1998[12]. En particulier, la variable de but figurait aussi sur le fichier de 1986 (mais pas sur le carnet) ; cette variable n'était pas directement remplie par les individus en 1986, elle était chiffrée par l'enquêteur en cas d'ambiguïté à partir des intitulés des activités. Pour l'enquête de 1998-1999, le nombre de modalités a été fixé à 4 : " personnel ou pour son ménage ", " professionnel ", " pour un autre ménage " et " associatif ".

Ainsi, à partir de ces variables annexes, une cinquantaine de règles de codage ont été initialisées. Elles étaient au nombre de 99 à la fin de l'enquête.

Exemple de règle de codage sur un exemple simple

Libellé écrit sur carnet

" **JE** PREPARE **UNE** PIZZA **POUR LES** ENFANTS ".

1ère étape: normalisation

" PREPARE PIZZA ENFANTS "

2ème étape : synonymisation

" PREPARE **ALIMENTS** ENFANTS "

3ème étape : application des règles avec variable de but

si but personnel	311
si but professionnel	211
si but pour un autre ménage	543
si but associatif	542
si but non renseigné	311

(arbitrairement choisi, car le plus fréquent)

[12] A savoir l'heure de début d'activité, l'heure de fin d'activité, la durée en minutes de l'activité et le statut professionnel de l'individu : salarié actif occupé, agriculteur actif occupé, autre indépendant actif occupé, étudiant ou élève ou autre (chômeur, retraité, personne au foyer ..).

4.5 L'utilisation de mots joker

L'étude des libellés en rejet et des libellés codés par Sicore vers différents codes (codés multiples) a permis de déceler un besoin d'utilisation de mots joker.

> *Par exemple, des prénoms apparaissent fréquemment dans les compléments d'activité : " dors sommeil " est suffisant pour classer l'activité, les informations supplémentaires risquant de parasiter le codage. Ainsi " Dors sommeil Pascal " ne serait pas codé.. En effet, la recherche Sicore est faite sur les quatre premiers mots, et dans ce cas, Sicore ne reconnaîtrait pas « Pascal ». La syntaxe de l'introduction du mot-joker est " Dors sommeil §§ ", signifiant que quelle que soit la valeur des deux derniers mots, il ne faut tenir compte que des deux premiers.*

Comme on peut l'imaginer, ce procédé radical nécessite une introduction circonspecte des jokers. Nous nous sommes limités aux libellés présentant le moins de risque, c'est-à-dire ceux possédant au minimum 2 mots pour qu'il n'y ait pas de contrôle de redondance sur un mot joker. Eventuellement, dans les cas rares où un seul mot était signifiant et sans ambiguïté, un mot joker doit être rajouté.

5. Les conséquences de l'utilisation de Sicore

5.1 Des modifications dans la présentation du carnet

Dès le premier test sur le terrain (en juin 96), les trois variables annexes citées au paragraphe 4.3 (" en présence de qui? ", le " lieu de l'occupation " et " votre activité est dans un but ... ") ont été présentées sur la page droite du carnet en face de chaque activité principale (cf. annexe n°3). Bien que ces variables ne servent pas systématiquement pour la codification de toutes les activités, on a demandé qu'elles figurent sur toutes les lignes du carnet, ce qui a été effectivement bien compris puisque sur le fichier final, on constate que le pourcentage de variables non renseignées est inférieur à 1%.

5.2 La modification de la nomenclature

La nomenclature utilisée lors des tests et au début de l'enquête a été réduite de 199 postes à 103.

En effet, de passer d'un codage manuel à un codage automatisé nous a obligés à établir des règles écrites, précises, qui ne laissent place à aucune ambiguïté et nous a amenés à agréger certains postes.

Prenons par exemple « manger »

Certains se limitent à ce seul descriptif, d'autres fournissent plus de détail par exemple « je mange un casse croûte ». Peut-être les premiers ont-ils aussi mangé un casse croûte, mais on ne le sait pas. Par contre on sait que dans les deux cas, il y a eu l'action de « manger ». Cela explique que dans la nomenclature de 1998, dans le doute, nous n'avons pas essayé de garder la nature du repas (casse croûte, collation, café, sandwich...). Par contre, nous avons gardé, le lieu du repas et avec qui était pris ce repas, informations qui pouvaient être connues grâce aux variables annexes. Les postes repas sont ainsi passés de 20 à 4.

De même « dormir »,

le poste " sommeil " est devenu un poste global agrégeant les siestes, et le repos.

De même le poste « transport »

nous n'avons plus distingué les différents modes de transport (à pied, en voiture, ...), réduisant le nombre de postes " trajets de 15 à 2. Il aurait fallu ajouter une variable annexe supplémentaire en cas de transport ce qui nous semblait alourdir le carnet inutilement. En effet, sur les carnets de 1986, on a constaté que le mode de transport qui devait figurer en clair, avait souvent été oublié. D'ailleurs, peu d'exploitations de 1986 ont été faites à ce niveau de détail.

Par contre d'autres décisions ont été prises indépendamment du fait qu'on utilise une codification automatique ou non.

Par exemple,

nous avons supprimé les codes qui étaient apparus en 1986 associés à des activités effectuées par moins de 100 individus.

Avec ce nouvel environnement, 80% des 3 253 libellés issus du test méthodologique de juin 1996 ont été codés.

Lors de l'enquête, par contre, nous avons ajouté des codes nouveaux. Ceci a été rendu possible grâce à l'utilisation des variables annexes et à la fin de l'enquête, la nomenclature est passée de 103 à 139 postes.

En effet, grâce aux variables de but, nous avons pu dissocier certains postes ; en particulier les postes concernant les travaux ménagers ou les aides, si l'activité était faite « pour un autre ménage » ou « dans un but associatif ».

6. Au bilan

6.1 Une réussite dans la quantité et la qualité du codage

Nous avions décidé, au moins pour les premières vagues, étant donné le caractère saisonnier de la variable, qu'une étude approfondie des rejets serait faite pour améliorer l'environnement Sicore. Pour cela, à chaque vague, l'expert variable, à partir de la liste des biens codés, triée par libellés et codes, vérifiait le bon codage. De plus, à partir des rejets triés par libellés, il pouvait comprendre pourquoi le libellé était mal codé et agir sur l'environnement Sicore pour qu'il soit codé lors des vagues selon le dispositif décrit ci-dessus.

Les libellés d'activités de l'enquête se répartissaient comme suit :

Activité	Principale	Secondire	Total
Nombre de libellés	316 097	113 581	429 678
Pourcentage	73,6%	26,4%	100%

L'ensemble de l'enquête comportait **429 678 libellés à coder**, la plupart étant des activités principales.

Les caractéristiques de l'environnement utilisé lors du recodage final sont les suivantes :

FAB :	13 400
Synonymes	***2 388***
Règles logiques	99
nomenclature	139
variables annexes	9

A l'issue d'un recodage complet des 8 vagues grâce à cet environnement, le taux de codage automatique de l'ensemble des libellés de l'enquête a atteint plus de 90% avec une qualité excellente dépassant les 95%.

Activité	Principale	Secondaire	Total
Taux de codage automatique	89,9%	98,2%	92,2%

L'activité secondaire a été très facile à coder et très bien codée. En effet, elle se résume à quelques postes (" conversation ", " radio ", " télévision "), souvent notés tels quels dans le carnet et qui à eux trois représentent plus de 90% de l'occurrence des activités secondaires.

Le gain d'efficacité entre la première vague et la dernière vague n'est que de 2 points environ, passant de 90,4% à 92,2%. Ce gain a surtout été fait par des enrichissements à l'issue des 3 premières vagues.

Pour mieux comprendre le gain d'efficacité entre la première et la dernière vague:
Lors du dernier test de septembre 1997, nous avions atteint une efficacité de 81,6% mais il ne s'agissait que des activités principales qui sont plus difficiles à coder que les secondaires. Il ne faut pas comparer ce 81,6% avec l'efficacité globale de la 1ère vague de 90,4% (ensemble des activités primaires et secondaires).

On peut cependant estimer l'efficacité globale du dernier test à 86%, en tenant compte de la part respective des activités primaires et secondaires dans le volume total des activités à coder.

Le gain d'efficacité obtenu par l'enrichissement des bases de connaissances fait par l'expert variable, après le dernier test, et servant au codage de la 1ère vague pourrait ainsi être estimé environ à 4 points (différence entre 90,4% et 86%).

6.2 Une réussite en gain de temps manuel

On constate que le temps passé pour traiter ces rejets d'activité est de 1 694 heures pour environ 40 000 rejets au total. Si l'on ajoute le temps de saisie des libellés (environ 6 000 heures), on constate un gain de 22 000 heures de temps manuel par rapport aux 30 000 heures de 1986. Ceci dépasse de loin les espérances les plus optimistes du début d'enquête.

Cette comparaison avec 30 000 heures en 1986, au lieu des 45 000 annoncées au début de l'article est due au fait que le volume des libellés à traiter en 1985 a diminué d'un tiers en 1998. Les nombres de carnets exploités dans les deux enquêtes étaient pourtant équivalents mais les carnets en 1986 contenaient en moyenne 30 activités contre 20 en 1998.

Le gain de cette méthode ne s'arrête pas là. En effet, c'est la première fois que nous avons à disposition un fichier de libellés d'activités quotidiennes de cette importance sur lequel nous pouvons travailler (analyse textuelle...), afin d'harmoniser les comparaisons à l'avenir, et effectuer toutes sortes de recodifications possibles.

Un gain certain en temps manuel, certes, mais il faut tenir compte du temps de création, gestion, formation à l'environnement Sicore qu'on peut estimer globalement à environ 6 mois de travail à temps complet par un contrôleur de l'INSEE.

6.3 Des échanges facilités par une organisation centralisée

Pour l'enquête, une organisation entièrement centralisée a été mise en place[13]. Tous les documents de l'enquête étaient traités par le GSAS de la DR des Pays de la Loire. Une petite équipe était spécialisée dans la reprise des rejets de codage sur la base du volontariat.

Dans le cas de reprises jugées délicates, et/ou dans le cas où les renseignements affichés à l'écran de reprise n'étaient pas suffisants pour coder l'activité, chaque codeur pouvait, grâce au traitement par lots, revenir au dossier papier complet pour mieux situer l'activité dans l'ensemble de la journée ou consulter des variables contenues dans l'ensemble du dossier. Il en référait au responsable de l'équipe qui tranchait.

Des réunions fréquentes avaient lieu au sein de l'équipe pour mettre en commun les problèmes rencontrés qui étaient alors transmis à l'expert variable par le responsable. Cette mise en commun des problèmes rencontrés a deux avantages. Tout d'abord, elle permet de rendre la codification manuelle assez homogène. D'autre part, elle facilite le travail de l'expert variable qui n'a affaire qu'à un seul interlocuteur.

6.4 L'impact négatif sur les comparaisons temporelles

Le succès de la codification automatique de la variable activité sur cette enquête doit être nuancé car ce codage réussi dépend des règles établies et de leurs variables annexes choisies. Cet ensemble de choix rend plus difficiles les comparaisons avec les enquêtes précédentes codées manuellement.

En effet, d'une part les libellés des activités de l'enquête de 1986 et des précédentes n'ont pas été saisis, ce qui rend leur codage automatique par Sicore impossible ; et d'autre part, dans les éditions précédentes de l'enquête EdT, les variables annexes n'existaient pas sous la même forme que celle de l'édition de 1998-1999.

6.5 Le choix non neutre des variables annexes et des règles

Le codage automatique a l'avantage d'être homogène sur l'ensemble des libellés de l'enquête. En revanche, en 1986, les variables annexes étaient notées " en clair " sur

[13] Pour plus de renseignements sur l'enquête Emploi du Temps, on se reportera à l'annexe n°2

le carnet quand elles l'étaient. Et c'était le codeur qui interprétait l'ensemble des informations du carnet, et codait les activités ; ce qui aboutissait à des codes hétérogènes. Par ailleurs, nous avons pu constater, sur certains carnets de 1986, que le codage manuel de l'époque n'était pas exempt d'erreurs.

Par contre, en cas de mauvaise écriture des règles, l'erreur de codage automatique entraînée sera systématique, alors qu'en 1986, une erreur dépendait des codeurs manuels et était plus diluée dans l'ensemble des activités. Cependant, si lors de l'exploitation de l'enquête de 1998, on se rend compte d'une telle erreur, on peut la corriger systématiquement en corrigeant la règle, ce qui n'est pas le cas pour 1986.
L'enquête de 1998-1999 est actuellement en cours d'exploitation et nous nous rendons compte de l'impact du remplissage des variables annexes et de leur choix dans les règles qui leur sont associées dans la codification automatique.

> *Par exemple, nous avons regardé le temps de travail des assistantes maternelles dont la profession consiste à garder des enfants chez elles*[14] *. Lorsqu'elles faisaient leur ménage, ou leurs courses, elles ont eu plutôt tendance à mettre " but personnel ", ce qui est tout à fait normal, et à cause de cette variable annexe de but qui joue prioritairement dans nos règles, cette durée n'a pas été comptée en temps quotidien de travail, mais en temps domestique.*
>
> *Le temps de travail, professionnel est calculé par ailleurs dans l'enquête soit par le semainier de travail, soit par une question globale (quel est votre temps habituel de travail ?). Au vu des résultats, on s'est rendu compte que le temps de travail moyen des assistantes maternelles calculé par le carnet tout au long de la journée, est nettement inférieur aux deux autres temps de travail. En effet, quand on leur demande globalement combien de temps elles travaillent, elles on plutôt tendance à donner le temps où elles sont responsables de l'enfant, et non celui où elles s'occupent réellement de lui.*
>
> *Si on avait eu à disposition, lors de la codification automatique, une CS plus fine, permettant d'isoler cette profession, on aurait pu coder différemment le travail professionnel des assistantes maternelles en construisant des règles adaptées sachant que lorsque l'enfant était avec elles, c'était un temps professionnel et non un temps domestique et rendre la variable de présence prioritaire par rapport à celle de but.*
> *Nous aurions alors perdu le détail de leurs activités quotidiennes, qui auraient été considérées alors comme des activités secondaires. Mais où est la vérité ?*

[14] La CS actuellement dans le fichier n'isole pas les assistantes maternelles. C'est un travail de décryptage de l'activité professionnelle en clair qui a permis à un chercheur exploitant l'enquête de les repérer.

Ceci met en évidence un grand avantage de l'enquête emploi du temps de 1998 par rapport à celle de 1986 . Au cours des exploitations, grâce au fait que les libellés soient en clair, chacun peut, avec ses moyens « recodifier » certaines activités, selon ses propres règles de codage.

6.6 Le souhait du responsable de l'enquête EdT

Le souhait du responsable d'enquête serait que l'expert Sicore qui a un poste permanent, fasse vivre et grandir cet environnement Sicore pour permettre, pourquoi pas, des enquêtes Emploi du Temps plus fréquentes ou codifier des libellés d'activités quotidiennes dans d'autres enquêtes. En effet, l'enquête Emploi du Temps actuellement n'a lieu que tous les 10 ans, et il serait dommage que cette base de connaissance devienne caduque faute d'un expert variable, qui en général, n'est nommé que pour la durée d'une enquête.

II - SICORE et les produits de dépense dans l'enquête Budget de Famille 2000-2001

1. La variable produit et son support de collecte : le carnet

1.1 Le carnet de comptes

Le carnet de l'enquête BdF court sur 14 jours et non sur un jour. Tous les individus de plus de 14 ans du ménage interrogé doivent inscrire toutes les dépenses qu'ils effectuent au cours de ces 2 semaines de collecte dans ce carnet. L'enquêteur se déplace 3 fois chez le ménage, en particulier pour expliquer l'objectif et l'utilisation du carnet, vérifier le bon remplissage sur la période de collecte et le ramasser.

1.2 Une variable hétérogène et non structurée

Mais revenons sur la définition d'une **dépense** dans l'enquête Budget de Famille. Elle se décompose en 4 entités : **une quantité, un produit, un montant et un lieu d'achat** (exemple : 1 - place de théâtre - 250F - FNAC). Elle est aussi associée à un individu d'un ménage appartenant à l'échantillon initial, et elle est repérée dans le temps par un jour parmi les 14 d'une vague donnée (cf. annexe n°6).

Cet individu peut avoir effectué cet achat de 2 manières différentes qui seront consignées dans le carnet. Ou bien, il a réalisé un certain nombre d'achats (plus de 10 articles) dans un seul lieu et possède un ticket de caisse comportant toutes les informations sur ses dépenses ; il devra donc coller son ticket sur la partie gauche du carnet. Ou bien, s'il a effectué une dépense isolée, il devra l'inscrire sur la partie droite du carnet dans le " tableau de dépenses ".

- Lorsque la dépense est isolée, l'individu titulaire du carnet doit remplir les 4 champs figurant dans le tableau (partie droite du carnet) :

Quantité/unité	Nature de la dépense (produit)	Montant de la dépense	Type de magasin (lieu d'achat)
3 litres	*lait*	*18,00*	*épicerie*
1	*place de théâtre*	*250,00*	*FNAC*

- Dans le cas où il colle son ticket de caisse (achats dans un supermarché), il doit vérifier que son ticket comporte sensiblement les mêmes informations que celles du tableau de dépenses (non compris la quantité et/ou l'unité qui

sont des informations non obligatoires). Si cela n'est pas le cas, il doit ajouter l'information manquante afin d'identifier la dépense aussi bien que s'il l'avait notée directement dans le tableau de dépenses.

Le plus souvent, les tickets de caisse ne comportent pas des intitulés de produits aussi facilement identifiables que ceux figurant sur la partie " tableau de dépenses ". Par conséquent, il doit identifier le produit de manière claire en complétant de manière manuscrite chacun des produits incompréhensibles : c'est ce que l'on appelle **le complément de la dépense.**

Exemples :

- *Si le libellé du ticket de caisse est " 1 pack Yoplait 20F ", le consommateur devra tout d'abord inscrire le magasin s'il n'apparaît pas sur le ticket. Il devra d'autre part compléter le produit Yoplait par " yaourts " Les informations sur le produit qu'il note constituent des compléments de dépense. Lors de la saisie des carnets, elles seront saisies isolément de l'intitulé du produit tel qu'il est inscrit sur le ticket.*
- *Dans l'exemple de ticket de caisse ci-dessous, figurent à droite les compléments de dépense censés identifier les produits achetés*

INTERMARCHE				
GINI 10x20CL 6 1/1.95	18.70			
VOLVIC	11.70			
OUIL.AUTO MAG	169.00	→	→	***4 housses siège voiture***
JOUETS LOISIRS	25.80			
MIEL ACACIA	15.30			
COQUIGRAIN 65/70x6	6.40	→	→	***6 oeufs***
BATONNETS COTON	3.45			
RICORE	20.30			
FEMME	49.90	→	→	***1 slip femme***
GENIE GEL	11.80			
MAJESTY BOEUF	2.10	→	→	***1 boite alimentation chat***
BTE PYRENE.	21.95			
TOTAL	***365.35***			

Il est clair que le produit intitulé " FEMME " est difficilement identifiable s'il n'est pas associé à un complément d'information.

Ainsi, les champs de saisie pour la **partie tickets de caisse** sont les suivants :

Nature de la dépense	Complément de la dépense	Montant de la dépense	Type de magasin
1 pack Yoplait	*Yaourts*	*20,00*	*Supermarché*

tandis que ceux de la partie du **tableau de dépenses** sont les suivants :

Quantité/unité	Nature de la dépense	Montant de la dépense	Type de magasin
3 packs	*lait*	*18,00*	*Epicerie*

Ainsi, les informations collectées pour un produit donné sont hétérogènes car elles proviennent de 2 supports différents : ticket de caisse ou tableau de dépenses.

D'autre part, il suffit d'examiner plusieurs tickets de caisse provenant de commerces différents pour s'apercevoir que l'information pour un même produit est non structurée contrairement à la variable activité. Les produits issus des tickets de caisse peuvent être des marques de produit, des chiffres, des abréviations et/ou des codes. Nous verrons dans le paragraphe 6 les différents problèmes rencontrés face à un libellé aussi hétérogène.

1.3 Les caractéristiques de la variable et leurs évolutions dans le temps

La saisie des carnets de l'enquête BdF précédente a permis de montrer que déjà en 1994-1995, les dépenses des ménages inscrites dans ces carnets provenaient majoritairement de tickets de caisse (à 56% environ). D'autre part, le complément de dépense était présent pour un tiers des dépenses provenant des tickets de caisse. Enfin, le nombre de dépenses par carnet représente un volume considérable. En novembre 1998, lors du premier test terrain, la tendance s'était accentuée.

Fichier	Carnets 1994-95	Test1[15] nov. 1998	Test2 juin 1999
Nombre de dépenses par carnet	57	65	89
Pourcentage de dépenses provenant de tickets de caisse	55,75%	60,9 %	63,01%
Pourcentage de dépenses issues de tickets de caisse comportant un complément	34,66%	37,22%	32,11%

[15] Le premier test a eu lieu en novembre 1998 dans les DR de Lille, Lyon et Nancy. Le deuxième test s'est déroulé en juin 1999 dans les DR de Lyon, Orléans et Rennes

Ces caractéristiques ont sans doute des conséquences sur les taux de codification. La présence d'un complément d'information pour un produit provenant d'un ticket de caisse aura un rôle important autant du point de vue de la codification automatique que celui de la reprise manuelle des non-codés automatiquement.

2 La nomenclature de la variable produits

2.1 La nomenclature initiale de l'enquête de 1994-95

La nomenclature officielle en 1994-95 se divisait en 916 postes, écrits sur 5 positions, emboîtés et divisés en 9 grandes fonctions :
Cependant, toutes les fonctions ne contenaient pas le même nombre de postes : l'alimentation était très nettement majoritaire sur les autres : les postes de la fonction " Produits alimentaires, boissons et tabac " représentaient 29% de l'ensemble des postes.

2.2 Le changement de nomenclature en janvier 1999

En 1997[16], Eurostat avait recommandé, dans des soucis de comparaisons internationales, une nomenclature de produits pour l'enquête BdF, la COICOP-HBS[17]. Celle-ci comprend 13 grandes fonctions, s'écrit sur 5 positions et comprend des postes emboîtés comme ceux de la nomenclature de 1994-1995. Néanmoins, afin de garder une continuité de nomenclature avec les précédentes éditions de BdF, nous avons rajouté une position et sommes arrivées à 6 positions. Les 13 fonctions de la nomenclature européenne sont celles-ci :

[16] cf. bibliographie

[17] Classification Of Individual COnsumption by Purpose - Household Budget Survey

Numéro de la fonction	Intitulé de la fonction	Exemples de produit
01	Produits alimentaires et boissons non alcoolisées	Pain, yaourts, jus d'orange
02	**Boissons alcoolisées, tabac et stupéfiants**	**Cigarettes, champagne**
03	Articles d'habillement et chaussures	Pantalon pour homme, jupe pour fillette, soutien-gorge, tissu, cravate, espadrilles pour garçon, **pressing pour les vêtements**
04	Logement, eau, électricité, gaz et autres combustibles	Loyer, GDF, charbon, **papier peint**
05	Ameublement, équipement ménager en entretien courant de la maison	Lit, sonnette, cafetière électrique, voilage, assiette, piles, savon, **pressing pour le linge d'ameublement, baby-sitter**
06	Santé	sirop contre la toux, sparadrap, lunettes, médecin, frais d'hospitalisation
07	Transports	Voiture, vélo, pneus, vidange, huile moteur, péage, billet de train, carte orange,
08	**Communications**	**Facture de téléphone, timbres**
09	Loisirs et culture	Télévision, disques, sac de couchage, planche à voile, monopoly, aquarium, bulbe de tulipe, flipper, cinéma, bowling, vacances, journaux, **agrafes, stylo**
10	**Enseignement**	**Frais de scolarité**
11	**Hôtels cafés et restaurants**	**Hôtel, repas, café au comptoir**
12	Autres biens et services	**Coton**, nourrice agréée, coiffeur, déodorant, bouillotte, réveil, cartable, **assurances**
13	Postes hors champ de la comptabilité nationale	Impôts, toiture, cadeaux, remboursement de prêts, **permis de chasse**

Remarque : les exemples de produits en gras de ce tableau représentent des produits ayant changé de fonction par rapport à la nomenclature de 1994-95.

Les exemples dans la dernière colonne de ce tableau prouvent que le changement par rapport à la nomenclature n'était pas mineur.

3. Pourquoi utiliser Sicore pour coder le produit ?

Pour l'enquête précédente de 1994-1995, le codage des produits figurant dans les carnets était centralisé au CNE de Toulouse. Il pesait aussi très lourd en heures manuelles de codage, comme le codage de l'enquête EdT de 1986.
Pour les mêmes raisons que celles évoquées dans le paragraphe I- 3 pour l'enquête EdT, il fallait réduire ces charges. La solution d'utiliser Sicore ayant été prise pour l'édition de 1998-99 de l'enquête Emploi du Temps, il était naturel de l'adopter aussi pour l'enquête Budget de Famille.

L'examen du codage manuel de 1994-1995 nous poussait aussi dans cette direction afin de rendre plus homogène le codage.

> *Par exemple, en 1994-95, le « savon » pouvait être codé manuellement dans la fonction 4 Equipement du logement ou dans la fonction 8 Autres biens et Services s'il se présentait sous la forme de gel douche ou pas.*

Pour l'édition 2000-2001 de l'enquête, et pour la 1ère fois depuis que l'enquête BdF existe, l'outil de codification automatique Sicore est utilisé pour coder, entre autres[18], les produits achetés et inscrits dans les carnets.

4. La constitution des connaissances

4.1 Le premier Fichier d'apprentissage brut des produits

Pour utiliser Sicore, il faut constituer en particulier un fichier d'apprentissage (Fab). Là encore, sur l'enquête BdF tout comme sur l'enquête EdT, aucun fichier n'existait. En effet, les dépenses des carnets de l'enquête BdF précédente n'avaient pas été saisies.

En 1994-95, l'équipe du Centre National d'Exploitation de Toulouse où tous les carnets de l'enquête étaient centralisés codifiait les dépenses figurant sur les carnets et recopiait les codes choisis et les montants associés sur des bordereaux. Ceux-ci étaient ensuite pris en charge par l'atelier de saisie sur place. Il n'y avait donc pas de fichier de libellés de dépenses prêt lorsque la décision d'innover avec Sicore a été prise à l'été 1998.

[18] D'autres variables sont en outre codées par Sicore : tout d'abord, la CSP au sein du tronc commun ; ensuite le type de magasin dans le carnet Cette dernière information pose nettement moins de problèmes de codification que le produit de la dépense ; nous n'en parlerons donc pas ici pour nous focaliser sur le produit.

A l'été 1998, l'expert variable[19] n'étant pas nommé, c'est l'expert Sicore[20] et le responsable d'enquête qui ont constitué les bases de connaissances de la variable produit dans leurs premières versions. A ces libellés de postes de nomenclature, ont été rajoutés des exemples similaires à ceux de la colonne du tableau de la nomenclature COICOP-HBS (cf. paragraphe II - 2.2) et des libellés de marques de produit issus d'un site de commerce électronique. A ce stade, **le Fab comptait 3 179 lignes**.

4.2 L'amélioration standard des connaissances pour la variable produit

Pour la variable produit, la boucle Sicore a permis l'enrichissement des connaissances suivant plusieurs axes à partir de trois sources d'informations : les carnets de l'enquête de 1994-1995 et ceux des deux tests terrain :

- Le choix des caractères blancs ou vides

- Le choix, l'intégration, la mise en ordre et l'élimination de synonymes

  On dénombre deux types de synonymes. Tout d'abord, on trouve les synonymes " classiques ".

  Par exemple,
  " BOCAL " = " CONSERVE ",
  " RIDEAU " = " VOILAGE ".

  Ensuite, viennent les synonymes créant de nouvelles expressions et aboutissant à une réduction du nombre de mots au sein du libellé normalisé.

  Par exemple,
  " HARICOT VERT " = " HARVER ",
  " FROMAGE BLANC " = " FROMBLAN ",
  " SOUS VETEMENT " = " SOUSVET ".
  Ces expressions ne correspondent pas forcément aux abréviations rencontrées sur les tickets de caisse.

- Le choix des mots vides

  Par exemple, toutes les quantités sont éliminées lors de la phase de normalisation : " GRAMME ", " PACK ", " LITRE ", " LOT ", ...

[19] Ce n'est qu'à partir de janvier 1999 qu'*Hélène Fréchou* a été nommée expert variable mais elle est partie en août 1999. De septembre à novembre 1999 , *Catherine Taché* a joué aussi ce rôle. Enfin, depuis février 2000, *Martine Legay* est l'expert variable attitré.

[20] En l'occurrence, Frédérique Deschamps

Certains mots sont éliminés mais cette élimination a entraîné d'autres modifications des connaissances.

> *Par exemple, la définition de " PAIRE "en mot vide entraîne un codage automatique similaire des libellés " UNE PAIRE DE CHAUSSONS " en tant que pantoufles et " CHAUSSONS " en tant que pâtisserie. Si on veut continuer à éliminer le mot " PAIRE ", il faut ajouter un synonyme au bon endroit dans la liste des synonymes : " PAIRE CHAUSSON " = " PAIRCHAUSSON ".*

- Les paramètres de l'apprentissage
Ces premiers codages prenaient en compte 4 mots de 12 caractères chacun après normalisation.

- L'extension du nombre de lignes au sein du Fab constitue l'essentiel du travail de constitution des connaissances.

Sicore a été lancé sur les différents fichiers disponibles, soit issus de la saisie de carnets de 1994-1995, soit issus de la saisie des carnets des deux tests.

- Les fichiers de libellés de produits associés à un code de référence (manuel) ont été traités de trois manières différentes selon le code automatique fourni éventuellement par Sicore.

  ① Pour les codés conformes[21], nous avons éventuellement tiré un échantillon, puis nous avons étudié tous les libellés concernés. Dans les cas où le code automatique était erroné, nous avons corrigé les connaissances en conséquence.

  ② Nous avons intégrés les codés non conformes dans le FAB après arbitrage sur le code.

  ③ Pour les libellés non codés, c'est-à-dire sans code automatique valide fourni par Sicore (les erreurs de redondance plus les échecs de codage), étant donné le nombre de postes de la nomenclature de 1994-1995, nous avons appliqué une méthode de traitement **par mot-clef.** Cette méthode consiste, pour un poste précis de la nomenclature, à rechercher tous les libellés dont un des mots est en relation avec ce poste.

[21] Pour plus de renseignements sur Sicore et ses connaissances, on se reportera à l'annexe n°1.

Par exemple, pour le poste " saucisses fraîches, cuites ou fumées " ; les mots-clef cherchés dans ce fichier des non codés sont : " saucisse ", " toulouse ", " strasbourg ", " morteau ", " francfort ", " chipo ", " merguez ", " herta ", " knacki ", " crepinette ", " godiveau ".

Cette méthode permet de traiter rapidement une grande partie des libellés non codés se référant au poste choisi et évite la manipulation fastidieuse de la grande quantité de papier que représente la nomenclature. Elle présente deux inconvénients : d'une part, le nombre de postes de la nomenclature étant important, cette tâche est répétitive et nécessite de l'imagination ; d'autre part, elle a pour conséquence la récupération de libellés comportant un des mots-clef, mais sans aucun rapport avec le poste choisi.

Sur l'exemple précédent, nous avons récupéré le libellé suivant : " TICKET DE BUS POUR STRASBOURG " ou " LIVRE SUR TOULOUSE ".
Nous les avons alors codés sous le code ticket de bus ou livre.

- Pour les fichiers de libellés de produits sans code de référence (pas de code automatique, ni de code manuel de 1994-95), le traitement est plus long, car il n'y a pas de référence à la nomenclature autre que celle issue de Sicore.

① Pour les codés automatiquement par Sicore, nous avons vérifié que le code fourni par Sicore était juste. Le cas échéant, nous corrigions les bases de connaissances.

② Pour les non codés automatiquement par Sicore, deux méthodes ont été employées.
Au début, un codage manuel de ces différents produits a été réalisé et cette liste de produits associés à un code était simplifiée et intégrée au Fab. Puis, lassées du fait du volume de la nomenclature, la fréquence d'apparition des libellés de produits a été utilisée pour traiter ces libellés. Le Fab a été complété par ces libellés de produits, codés manuellement par nos soins, non existants jusqu'alors dans le Fab et dont la fréquence d'apparition était supérieure à 1.

4.3 L'utilisation optimale des informations auxiliaires

Des tests ont été menés sur les carnets saisis du test 1 de novembre 1998 afin de mesurer l'impact de la présence d'informations sur le codage automatique : le complément de dépense et la quantité.

Nous avons testé si ces informations étaient mieux placées en début ou en fin de libellés et si elles étaient nécessaires. Les résultats des multiples codages sont les suivants :

Type de dépenses	Total	Codés	Redondants	Non codés
avec le complément de dépense après le libellé de produit et la quantité avant	18 998	8 729 (45,95%)	4 736 (24,94%)	5 530 (29,11%)
avec le complément de dépense après le libellé de produit et sans la quantité	18 998	8 886 (46,77%)	4 711 (24,80%)	5 398 (28,41%)
avec le complément de dépense et la quantité avant le libellé de produit	18 998	8 241 (43,38%)	4 814 (25,35%)	5 940 (31,27%)
avec le complément de dépense avant le libellé de produit et sans la quantité	18 998	8 398 (44,20%)	4 789 (25,21%)	5 808 (30,57%)
avec le libellé seul, sans complément ni quantité	18 998	9 394 (49,45%)	4 458 (23,47%)	5 143 (27,07%)

➡ De ces tests de codage, il ressort que, si on intègre le complément de dépense, il vaut mieux qu'il soit placé après le libellé de la dépense. Quant à la quantité, le codage est meilleur lorsque le libellé ne contient pas de quantité.

4.4 Un nombre de mots plus adapté

Lors de l'étude réalisée sur les carnets saisis du test 1 de novembre 1998 (cf. paragraphe II - 4.3), nous avons aussi remarqué que les meilleurs taux de codage automatique étaient obtenus en ne prenant ni le complément de dépense ni la quantité (cf. tableau précédent). Ce résultat était surprenant : en général, des précisions permettent de mieux coder.

L'explication de ce problème vient du nombre de mots pris en compte dans les paramètres de codage Sicore : 4 dans ces codages. En effet, le complément d'information ne sert que quand le libellé du ticket de caisse est très court (sur 1 mot par exemple) ce qui est rare compte tenu des informations parasites qui font souvent

partie du libellé provenant d'un ticket de caisse, même si certaines sont éliminées par la normalisation.

Pour tenir compte du complément de dépense, nous avons alors réalisé des tests sur le nombre de mots du libellé normalisé comprenant le complément à prendre en compte. Jusqu'ici, nous prenions en compte un libellé de 4 mots, ce qui était insuffisant. Les résultats de ces tests montrent que considérer un libellé à 6 mots permet de coder plus de produits que prendre en compte 4 mots d'un libellé. En revanche, il ne sert à rien, a priori, de prendre en compte 8 mots.

Nombre de mots du libellé normalisé	Codés	Redondants	Echecs
4 mots	8 886 (46,77%)	4 711 (24,80%)	5 398 (28,41%)
6 mots	11 155 (58,72%)	2 524 (13,29%)	5 316 (27,958%)
8 mots	11 155 (58,72%)	2 524 (13,29%)	5 316 (27,958%)

➡ Au bilan, les codages prennent 6 mots dans le libellé issu de la normalisation.

4.5 La non-utilisation des variables annexes

Comme pour la variable activité, il a été question un temps de présenter une partie du carnet de dépenses (le tableau de dépenses) avec des variables annexes du type : destination de la dépense (pour l'individu, pour le ménage, pour la famille, pour d'autres personnes), lieu de dépense (grande surface, petite surface, autre cas).

Mais, compte tenu du comportement actuel des ménages à grouper de plus en plus leurs achats (par exemple sur le test 2 de juin 1999, le pourcentage de produits provenant de tickets de caisse s'élève à 63%), le problème des tickets de caisse demeurait.

Enfin, le fait d'avoir le lieu de dépense en variable précodée (en quelques modalités) comme le but de l'activité dans l'enquête EdT, limitait les possibilités d'exploitation et de collaboration avec les autres services divisions de l'INSEE : la division des prix à la consommation et les divisions Services et Commerces s'intéressent aux lieux d'achats. Ainsi, nous avons abandonné cette piste, mais nous gardons en tête l'idée des règles de codage (cf. paragraphe II - 8.2).

5. Les conséquences de l'utilisation de Sicore

5.1 Les modifications dans la présentation du carnet

La présentation des carnets s'est vue simplifiée par rapport à la mise en page des carnets de 1995 car nous avons tout d'abord essayé de faciliter le codage automatique des dépenses[22]. Ainsi, dans le carnet, le ménage n'inscrit que ses dépenses quotidiennes.

Ensuite, un seul type de carnet par rapport à l'enquête précédente a été utilisé : toutes les personnes du ménage ont un carnet du même type à remplir[23].

En novembre 1998, le test 1 avait aussi pour objectif de tester la présentation des carnets. En effet, les carnets utilisés étaient de deux types. Les premiers comportaient une colonne " quantité " isolée de la colonne " nature de la dépense " sur les pages concernant les dépenses isolées (tableau de dépense), tandis que les seconds fusionnaient ces deux informations. Au bilan, nous avons conservé un tableau de dépenses avec la colonne " quantité/unité " isolée.

Il a été décidé aussi de faciliter le plus possible le travail de remplissage du carnet par l'individu en lui laissant de la place pour inscrire ses dépenses sur le tableau de dépenses ou compléter l'information fournie par les tickets de caisse : le format des carnets a donc été changé par rapport à celui adopté en 1994-95 ; on ne peut plus réellement parler de carnet de dépenses, c'est davantage un cahier de dépenses.

Ainsi le tableau de dépenses du carnet comprend 4 champs :
- *quantité et unité,*
- *nature de la dépense,*
- *montant de la dépense,*
- *type de magasin.*

Et, lors de la formation des gestionnaires d'enquête qui forment eux-mêmes les enquêteurs, l'attention s'est portée plus particulièrement sur le complément de dépense à ajouter dans le cas de produits provenant de tickets de caisse.

[22] Deux types de tableaux ont été transférés sur CAPI ; ils concernaient l'autoconsommation et les repas pris à l'extérieur du domicile du titulaire du carnet et de ses enfants de moins de 14 ans (éventuels).

[23] En 1994-1995, la personne responsable de la majorité des dépenses du ménage devait remplir un carnet de type I comprenant, à la fois, un tableau de dépenses, une page pour les tickets de caisse, mais aussi le tableau sur l'autoconsommation du ménage et le tableau sur les repas pris à l'extérieur du domicile des personnes de moins de 14 ans (dispensées de carnet). Les autres personnes du ménage possédaient un carnet de type II contenant tout excepté le tableau décrivant l'autoconsommation du ménage et les repas pris à l'extérieur du domicile pour les personnes de moins de 14 ans.

5.2 L'identification difficile de certains produits

Sicore ne distingue pas les majuscules des minuscules et ne tient pas compte de la ponctuation. Ceci a pour conséquence que certains produits ne peuvent pas être codés de manière certaine ; des choix ont donc dû être faits.

> *Par exemple, " PATE " peut être codé vers plusieurs codes en 01 (pâtes Panzani, pâté en terrine ou pâte feuilletée) et vers le 09 (Pâté, pour chiens). On peut citer aussi " HACHE " qui peut être codé dans la fonction 09 (horticulture) ou 01 (viande).*

D'autre part, le fait que l'on n'utilise pas de variable annexe pour coder automatiquement les produits entraîne une identification plus délicate de certains produits.

> *Par exemple, le terme " COTON " peut correspondre à plusieurs produits : le coton à tricoter (03), le coton à démaquiller (12). Autre exemple, " CREME " peut être fraîche (01), Nivéa (12). " CARTOUCHE " fait référence aux cigarettes (02), à l'encre et à la chasse (09). De même, " HUILE " peut être alimentaire (01) ou pour la voiture (07).*

En revanche, un codeur aboutirait à arbitrer ces cas facilement à partir des informations auxiliaires (comme le lieu de dépense ou la quantité) disponible en clair pour lui. Des traitements spécifiques sont donc envisagés mais uniquement à la fin du traitement des carnets. A ce moment là, une règle de codage sera utilisée ; elle affinera le codage du produit à partir du magasin.

> *Par exemple, " COTON " associé au lieu de dépense " PHARMACIE " pourra être codé à la bonne place en 12.*

5.3 L'adaptation à la nomenclature

Il a été arbitré des codes pour des libellés ambigus pour la codification automatique par Sicore.

> *Par exemple, le libellé " PATES " sera codé comme les coquillettes Panzani et le libellé " PATE " comme les rillettes.*

D'autre part, pour les libellés généraux, ne permettant pas de déterminer les 6 chiffres du code, le caractère " * " a été inséré dans des codes. Ces cas nécessitent absolument la présence du complément de dépense pour être codés au plus juste. Sans ce dernier, on reste à un niveau général agrégé[24].

> *Par exemple, " BOUCHERIE " ne permet pas de déterminer quelle viande a été achetée ; on lui a donc attribué le code 0112**. De la même manière, a été créé le code à 6 étoiles " ****** " afin de coder des libellés ne contenant aucune information précise, comme " TICKET DE CAISSE ", " LIQUIDE ", ...*

Enfin, il arrive que certains libellés comprennent plusieurs produits de nature différente, et même appartenant à des fonctions différentes. Dans ce cas de multiplicité de produits, le plus grand nombre de chiffres en commun pour les codes est gardé et complété par des étoiles.

> *Par exemple, " FLEURS DENTIFRICE " sera codé également en ******. Mais, " OEUFS YAOURTS " est codé en 0114**.*
> *Lors de la future exploitation des fichiers, il faudra traiter ces cas.*

5.4 Le dédoublement du Fab des produits

L'examen attentif des produits issus des tickets de caisse met en évidence des libellés ayant une forme différente des libellés issus du tableau de dépense (cf : II - 6).

L'expert variable a alors décidé de dédoubler le Fab produits afin de tenir compte des spécificités des libellés des tickets de caisse :
- elle a supprimé du Fab des produits issus des tickets de caisse, les libellés de produits qui avaient très peu de chance d'être achetés dans des grandes surfaces
- elle a alors concentré ses efforts sur les produits alimentaires qui sont les produits les plus fréquents sur les tickets de caisse.

[24] Cf. le paragraphe 6.2

5.5 Une organisation des traitements des carnets complètement éclatée

Compte tenu du volume des carnets à saisir[25], tous les ateliers de saisie en exercice en mai 2000 ont dû travailler sur les carnets de l'enquête BdF 2000-2001. Etant au nombre de 13 contre 18 DR de collecte, des regroupements de DR ont été effectués selon la taille des GSAS.

Pour la reprise des carnets de dépenses, le volume de rejets à traiter[26] et l'organisation de la saisie ont eu des conséquences non négligeables sur l'organisation de la reprise qui est en fait, tout comme la saisie, bien différente de celle choisie pour EdT.

Nous avons pu maintenir le même nombre de sites de reprise : 13 DR réalisent ce travail grâce à un outil de reprise[27]. En revanche, le travail de codage manuel peut être fait par des équipes différentes de celles qui effectuent la collecte (les DEM) ou la saisie (GSAS).

Cette organisation ne facilite

-ni l'homogénéisation des codes manuels attribués aux produits rejetés par Sicore.

-ni le travail de l'expert variable qui doit jongler avec les 13 interlocuteurs différents au minimum[28].

[25] Cf. annexe n°5

[26] Ce volume de rejets à traiter a été calculé à partir d'une prévision de taux de codage relativement faible mais réaliste à l'époque (cf. annexe n°5)

[27] Cet outil de reprise a été développé par Anne-Marie Duval au CNI de Lille (cf. annexe n°5)

[28] En effet, dans certaines DR, le travail de reprise peut être effectué à la fois par la DEM et par le GSAS ce qui multiplie les interlocuteurs.

6. Les conséquences des caractéristiques des dépenses recueillies

L'examen des libellés provenant à la fois des tableaux de dépenses et des tickets de caisse nous a permis de déceler des problèmes.

6.1 L'orthographe des libellés et les abréviations

- Des fautes d'orthographe sont fréquentes dans la partie tableau de dépenses et gênent la codification automatique quand elles n'aboutissent pas à un échec du codage.

Par exemple, les termes " YAOURT ", " BEEFSTEACK " ont de nombreuses variantes.

- Les tickets de caisse décrivent souvent les produits grâce à des abréviations et des codes internes au magasin.

Par exemple, nous avons déniché les libellés suivants assortis de leur complément de dépense :

" P ELEA.1.2.ECR LAIT "
" 2X250ML SH2/1 SEC/ SHAMPOOING "

De plus, " COQUI " peut amener aux produits suivants : des coquillettes ou une coquille de poisson ; et " BOUCHE " peut renvoyer à un article de boucherie ou à une bouchée à la reine.
Les abréviations de " YAOURT ", " BEEFSTEACK , " POMMES DE TERRE " sont aussi nombreuses.

➡ La majorité des fautes d'orthographe et des abréviations, en tous cas les recensées, ont été rajoutées sous forme de synonymes. Les codes internes ont eux été intégrés dans la liste des mots vides relative aux connaissances.

6.2 La dilution de l'information

.

Pour certains libellés, l'information principale qui permettrait un codage automatique facilement est diluée par des informations parasites, comme la couleur, la marque, la quantité, le type de produit, ... Ces problèmes sont particulièrement vrais pour les libellés issus des tickets de caisse. En effet, les tickets de caisse présentent souvent l'inconvénient de décrire le produit en une multitude de mots à cause de la présence du complément de dépense et de la quantité et souvent de la marque.

Par exemple,

- nous avons trouvé les libellés suivants assortis de leur éventuel complément de dépense :

" TENNIS REVERS CUP'S CHAUSSETTES FEMME "
" GRESSINS ALICE BAGUETTES POUR CARPACCIO "
" MARTI RIZ 1/2 LG ETUVE "
" PAIN 24 TRS FIN PAIN DE MIE "
" 1 ER PRIX1/2 ECREME LAIT "

- nous avons aussi trouvé les libellés suivants sans complément de dépense :

" P JEAN GARN BOUCHEE REINE "
" BA FRUIT SOLEIL 6X1 YAOUR "

" FL 750ML PAI VAISSELLE "

➡ L'augmentation du nombre de mots du libellé normalisé pris en compte lors du codage a permis de régler en partie ce problème.

6.3 L'absence d'information pertinente

Au contraire du cas précédent, l'information principale qui permettrait un codage automatique facilement n'apparaît pas dans le libellé faute d'un complément de dépense. Une fois de plus, ces problèmes sont particulièrement vrais pour les libellés issus des tickets de caisse.

Par exemple, les libellés suivants sont trouvés :

" NON ALIMENTAIRE "
" FR ET LEGU "
" FRIT M "

➡ L'acceptation de codes agrémentés de " * " a permis de régler ce problème.

6.4 Les non-dépenses présentes dans les tickets de caisse

Les tickets de caisse présentent des lignes qui ne correspondent pas à des dépenses stricto sensu : les annulations de produit précédemment saisi, les remises et ristournes, les consignes, les bons de réduction, les bons d'achat, ...

Les libellés suivants ont été trouvés :

AUBERGINE ANNULATION ARTICLE
BON REDUCT.
REMISE CARTE PASS FRAISES 250G

➡ Un code particulier a été attribué à ces différents libellés : 999999. Lors de l'exploitation, nous saurons ainsi qu'il faudra soit retrancher le montant associé, soit l'éliminer purement et simplement.

7. Au Bilan

7.1 Les résultats de codage au cours de la préparation des connaissances

① Chronologiquement, la préparation des connaissances a commencé par l'exploitation des carnets de 1994-1995. A cette époque, Sicore prenait en compte les 4 premiers mots du libellé normalisé. Avant l'utilisation des libellés du test 1, les résultats de codage de Sicore sur un fichier de référence[29] étaient les suivants :

Résultats du codage automatique du fichier de référence	Nombre de lignes	Pourcentage
Codés	20 395	**53,21%**
Non codés	17 931	46,79%
Total	38 326	100%

② Avec la même version des connaissances qui a fourni ces résultats, voici les résultats de codage des produits du test 1 de novembre 1998 :

Résultats du codage automatique des produits de novembre 1998	Nombre de lignes	Pourcentage
Codés	8 767	46,15%
Non codés	10 231	53,85%
Total	18 998	100%

Ces résultats de codage sont moins bons que ceux obtenus précédemment sur le fichier de référence des produits de 1994-95. Cela s'explique par les nouveaux produits apparus depuis cette époque et l'augmentation du pourcentage des libellés provenant des tickets de caisse.

③ Par la suite, les libellés du test 1 ayant servi à l'enrichissement des bases de connaissances et à prendre la décision d'utiliser 6 mots dans le libellé normalisé, nous avons recodé le fichier de référence. Ce genre d'opérations nous permet de mesurer les progrès engendrés par l'exploitation des libellés du test 1. Les résultats de codage de ce fichier ont donc été tout à fait encourageants ; nous avons gagné plus de 10 points dans le codage (53,21% contre 66,33%) :

[29] Ce fichier de 38 326 lignes était issu des carnets de 1994-95. Il n'a pas été utilisé pour enrichir les connaissances de Sicore. Il nous a servi de référence.

Résultats du codage automatique du fichier de référence	Nombre de lignes	Pourcentage
Codés	25 421	**66,33%**
Non codés	12 905	33,67%
Total	38 326	100%

④ Cette même version des connaissances a servi aussi à coder les produits du test 2 de juin 1999.

Résultats du codage automatique des produits de juin 1999	Nombre de lignes	Pourcentage
Codés	9 118	**60,24%**
Non codés	6 017	39,76%
Total	15 135	100%

Même si les taux de codage étaient inférieurs à ceux obtenus sur le fichier de référence (66,33% contre 60,24%), ces résultats ont été plutôt satisfaisants compte tenu du fait que le dernier enrichissement du Fab a été réalisé à partir de produits de novembre 1998, c'est à dire de certains produits saisonniers spécifiques à l'automne. Or, le test 2 s'étant déroulé au printemps, d'autres produits spécifiques ont pu apparaître sans avoir pu être codés.

⑤ Un sous-fichier du test 2[30] a permis également de fournir les premiers résultats de codage après la décision de dédoubler le Fab des produits. Ainsi, à cette époque, nous avons pu constater l'écart de codage égal à 25 points entre produits issus de tickets de caisse et produits issus des tableaux de dépense.

	produits des tickets de caisse		produits des tableaux	
Résultat du codage automatique des produits de juin 1999	**Nombre de lignes**	**%**	**Nombre de lignes**	**%**
Codés	2 896	**51,96%**	3 098	**77,28%**
Non codés	2 678	48,04%	911	22,72%
Total	5 574	100%	5 574	100%

[30] Ce sous fichier est constitué des produits recueillis dans la DR de Rhône Alpes lors du test 2 de juin 1999. Deux autres DR ont participé à ce test.

7.2 - Le premier environnement utilisé pour l'enquête en juin 2000

L'enquête a démarré sur le terrain en mai 2000. En juin 2000, les premiers codages ont été effectués. Les caractéristiques générales de ce premier environnement sont les suivantes :

Environnement	TABLEAU DE DEPENSES	TICKET DE CAISSE
Nombre de mots	6	6
Nombre de lettres par mots	12	12
FAB :	37 394 lignes	31 299 lignes
Nombre de codes comportant au moins une *	462	80
Nombre de synonymes	1 193	1 193

Voici la répartition des libellés de produits par grande fonction de la nomenclature :

Environnement	TABLEAU DE DEPENSES	TICKET de CAISSE
Fonction	nombre de libellés	nombre de libellés
01	22 696 (60,69%)	20 998 (67,09%)
02	1 398	1 187
03	1 321	1 204
04	418	274
05	2 908	2 585
06	491	146
07	641	206
08	129	16
09	3 551	2 638
10	40	0
11	576	6
12	2 253	1 786
13	573	147
99	37	26
Total	37 394	31 299

Il est logique de voir que le nombre de libellés dans le Fab produits issus des tickets de caisse dans le domaine de l'enseignement est nul et que la part de l'alimentaire (fonction 01) dans ce Fab est la plus importante (67,09%).

7.3 Les premières vagues de l'enquête

7.3.1 Les caractéristiques des libellés recueillis au début de l'enquête

Par rapport aux caractéristiques des carnets du test 2 de juin 1999, le pourcentage de produits des vagues 1 et 2[31] provenant des tickets de caisse a légèrement diminué : 63,01% contre 61,13 % en 2000 ce qui peut être une bonne chose pour le taux de codage automatique puisque la faiblesse du codage demeure sur les produits des tickets de caisse.

	Nombre de carnets	Nombre de produits	Provenance du produit	Nombre de produits
Vague 1	2 002	137 740	Ticket	83 981 (60,97%)
			Manuscrit	53 759 (39,03%)
Vague 2	1 785	129 309	Ticket	79 322 (61,34%)
			Manuscrit	49 987 (38,66%)
Total	3 787	267 129	Ticket	163 303 (61,13%)
			Manuscrit	103 746 (38,87%)

D'autre part, comme lors des formations des gestionnaires d'enquête, nous avions insisté sur l'importance d'un complément de dépense dans le cas d'un produit issu des tickets de caisse, nous avons voulu connaître ce nombre de produits issus des tickets de caisse qui comportaient un complément de dépense. Dans les carnets du test 2, le complément existait dans seulement 32,11% des cas. Et nous avons été agréablement surpris avec un taux au-dessus de 40 % pour les deux premières vagues de l'enquête.

	Nombre de produits issus des tickets de caisse	Existence de complément de dépense
vague 1	83 981	38 379 (45,70%)
vague 2	79 322	33 111 (41,74%)
Total	163 303	71 490 (43,78%)

7.3.2 L'enrichissement des connaissances en cours d'enquête

Comme pour l'enquête EdT, l'expert variable enrichit les Fab produits au fur et à mesure des vagues afin de bien améliorer le taux de codage automatique et par

[31] La vague 1 couvrait la période du 9 mai au 18 juin 2000 tandis que la vague 2 s'étalait du 19 juin au 30 juillet 2000 en métropole.

conséquent de diminuer le volume des produits à coder manuellement dans les DR assurant la reprise[32]. Elle vérifie aussi le codage automatique sur un échantillon de libellés.

L'enrichissement des fichiers de référence porte sur :

- l'aspect quantitatif : les libellés non reconnus (échecs) ou reconnus partiellement (redondants) sont isolés puis triés par fréquence d'apparition. Les libellés ou type de libellés les plus fréquents sont traités en priorité, soit en ajout simple, soit en synonymie puis ajoutés.
- l'aspect qualitatif : les libellés codés automatiquement sont vérifiés, " individuellement " s'il s'agit de libellés à haute fréquence d'apparition, par tirage d'échantillon parmi l'ensemble des codés et ponctuellement par poste de la nomenclature.

Par ailleurs, des contrôles sont opérés par l'expert variable lorsque les DR signalent des anomalies dans la codification automatique.

Après tous ces travaux, au 30 octobre 2000, quatre versions des environnements produits tickets de caisse ont été livrés et ont permis d'avoir un taux de codification automatique en augmentation.

En effet, sur la **vague 1**, le taux de codage automatique est de **69,22%** quel que soit l'environnement et la provenance de la dépense alors que le test 2 de juin 1999 était codé seulement à 63,26%. Sur la **vague 2,** ce taux passe au-dessus de 70% : **71,67.**

Résultat du codage automatique PRODUITS	**Nombre de produits VAGUE 1**	**% VAGUE 1**	**Nombre de produits VAGUE 2**	**% VAGUE 2**
Codés	95 345	**69,22%**	92 674	**71,67%**
Non codés	42 395	30,78%	36 635	28,33%
Total	137 740	100,00%	129 309	100,00%

7.3.3 Les produits des tickets de caisse de plus en plus codés

La mise à disposition de nouveaux environnements portant sur les libellés des produits issus des tickets de caisse sur les premières vagues de l'enquête est tout à fait payante. En effet, ce sont sur les libellés de tickets de caisse qu'ont porté tous les efforts de l'expert variable. On passe ainsi de 61,58% sur la vague 1 à 65,58% sur la vague 2 c'est-à-dire un gain de 4 points.

[32] Pour plus de renseignements sur l'enquête BdF, on se reportera à l'annexe n°5

VAGUE 1	produits des tickets de caisse		produits isolés	
Résultat du codage automatique	**Nombre de produits**	**%**	**Nombre de produits**	**%**
Codés	51 715	**61,58%**	43 630	**81,16%**
Redondants	10 044	11,96%	2 721	5,06%
Non codés	22 222	26,46%	7 408	13,78%
Total	83 981	100 %	53 759	100 %

VAGUE 2	produits des tickets de caisse		produits isolés	
Résultat du codage automatique	**Nombre de produits**	**%**	**Nombre de produits**	**%**
Codés	52 016	**65,58%**	40 658	**81,34%**
Redondants	9 204	11,60%	2 402	4,81%
Non codés	18 102	22,82%	6 927	13,86%
Total	79 322	100 %	49 987	100 %

Nous sommes plutôt satisfaites des résultats obtenus sur les deux premières vagues de l'enquête par rapport au taux de codification des tests mais le travail d'enrichissement du Fab produits est quasi quotidien et par conséquent fastidieux.

7.3.4 La découverte de nouveaux problèmes liés à la collecte des tickets de caisse

Nous nous sommes aperçues en examinant les carnets des vagues 1 et 2 d'un phénomène qui gênait (voire même empêchait) le codage automatique des produits et que nous n'avions pas du tout rencontré lors des tests.
Sur la partie ticket de caisse des carnets, nous avions demandé que soit ajouté un complément de dépense pertinent pour coder plus facilement le produit. Or compte tenu du fait que sur la partie tableau de dépense, la colonne " quantité / unité " est clairement séparée du reste, nous récupérons assez fréquemment une quantité dans le complément de dépense du ticket de caisse.

Nous avons donc essayé de rappeler aux gestionnaires d'enquête qui eux-mêmes transmettent cette information aux enquêteurs sur le terrain que le complément de dépense ne devait pas contenir uniquement la quantité et unité du produit. L'information ajoutée devait être " pertinente ".

Note : dans les Directions Régionales où c'est la même équipe qui forme les enquêteurs et qui réalise la reprise des non codés, le phénomène est moins fréquent. En effet, elle comprend mieux l'idée de pertinence pour un complément de dépenses.

8. Les prévisions

8.1 Pour les vagues suivantes

Des facteurs de variation dus à la variable produit peuvent laisser présager des disparités fortes de codage sur l'ensemble de l'enquête.

Tout d'abord, le marché de la grande consommation est tel que des nouveaux produits apparaissent chaque jour, inscrits le plus souvent dans les tickets de caisse. De ce point de vue, il semble que le travail de suivi des bases de connaissances devra se poursuivre tout au long de l'enquête.

Ensuite, comme nous l'avons déjà dit, la variable a un fort caractère saisonnier. Or, les deux tests ont porté sur de courtes périodes ne couvrant pas l'année entière. Certes, nous pouvons espérer que les carnets saisis de 1994-1995 ont couvert une bonne partie de l'année, mais ils étaient peu nombreux et commencent à vieillir un peu. En particulier, les périodes comme celles des fêtes de fin d'année et des vacances d'hiver représentent des périodes de consommation spécifiques non rencontrées ou presque jusqu'ici.

Enfin, les tests ont eu lieu sur une petite partie du territoire métropolitain[33]. On peut s'attendre à ce que des produits régionaux spécifiques, comme ceux de la Corse ou d'Alsace-Lorraine, ne soient pas pris en compte dès maintenant : du travail reste à fournir pour ces cas-là.

8.2 A la fin de l'enquête

A la fin de l'enquête, une règle de codage sera utilisée ; elle affinera le codage du produit à partir du magasin.

> *Par exemple, un café acheté dans une épicerie sous forme de paquet et un café payé au comptoir sont actuellement codés automatiquement avec le même code. Or, la nomenclature des produits (COICOP-HBS) donne 2 codes différents.*

De plus, nous effectuerons aussi un recodage complet avec la dernière version des bases de connaissance de tous les produits afin d'harmoniser les codes des produits. Nous traiterons, aussi, les codes 999999 correspondant aux non dépenses (cf. paragraphe II - 6.4).

[33] L'enquête a aussi lieu dans les DOM avec les mêmes Fab complétés par quelques produits plus locaux.

Conclusion : le mot de l'ex-expert Sicore

Ce papier présente la mise en place d'un dispositif d'enquête sur deux enquêtes assez lourdes. En effet, utiliser la codification automatique au sein d'une enquête est loin d'être neutre, même si le gain (essentiellement en heures manuelles) est réel. Nous avons donc décrit ici les grandes étapes de la mise en place de la codification automatique sur deux variables nouvelles : les activités quotidiennes et les produits de dépense, ainsi que les principaux problèmes concrets rencontrés.

En ce qui concerne les activités quotidiennes, une bonne partie du travail préparatoire à l'enquête sur les bases de connaissances a porté sur **la construction des règles de codage**. En effet, pour l'enquête Emploi du Temps, la nomenclature et la structure du recueil des données (le carnet) se prêtaient bien à l'utilisation des règles de codage. Evidemment, le travail a aussi consisté à constituer et enrichir le fichier d'apprentissage. Toutefois, le taux de codification automatique a dès le début été bon : la maintenance des bases de connaissances en cours d'enquête a été relativement légère[34]. Au final, le taux de codification automatique a atteint 92,2% et on a constaté une excellente qualité du codage final.

En revanche, pour les produits, le travail préparatoire à l'enquête BdF pour construire les bases de connaissances a été relativement colossal et fastidieux, essentiellement parce que, pour une même variable, nous traitions une information hétérogène : les produits isolés transcrits manuellement par l'enquêté et les produits issus des tickets de caisse collés sur le carnet. Et, nous avons vu que l'information recueillie via les tickets de caisse était toute différente et nécessitait un traitement à part. Ainsi, deux bases de connaissances sont nécessaires pour coder une seule et même variable. Tout cela a entraîné (en plus de la vacance du poste d'expert variable pendant un laps de temps non négligeable) une préparation insuffisante avant le démarrage de l'enquête. Finalement, les taux de codification restent bas (70%) et le travail sur les bases de connaissances toujours d'actualité, étant donné que l'enquête est en cours.

Outre la présentation de méthodes pouvant servir à la mise en place de la codification automatique d'une nouvelle variable par Sicore, ce papier montre les difficultés d'une telle opération : le résultat de la codification ne dépend pas uniquement de la structure de l'enquête sur laquelle elle est mise en place ; elle dépend également de la variable intrinsèque qui est codée (nomenclature, structure de l'information recueillie, qualité du recueil de l'information).

[34] Dans la mesure où le travail de maintenance de bases de connaissances relatives à une variable via Sicore peut être léger.

Loin de prouver que l'utilisation de Sicore repose sur quelques principes généraux lors de l'application, nous avons montré que nous ne pouvons pas préconiser des conseils universels à appliquer lors de la construction et de l'amélioration des bases de connaissances. En réalité, les seules recommandations que l'on pourrait donner seraient d'être tenace et de rester patient ...

Bibliographie

1 - Sur la codification automatique en général

LORIGNY, J. (1988). QUID, une méthode générale de chiffrement automatique. *Survey methodology, décembre 1988,* Vol.14, n°2, pp.289-298.

LYBERG L., DEAN P. (1992) Automated coding of survey responses : an international review Working Paper, *Conférence des Statisticiens Européens, Work Session on Data Editing,* Washington, Mars 1992

WENZLOWSKI, M.J. (1988) ACTR - A generalized automatic coding system *Survey Methodology, décembre 1988,* Vol.14, n°2, pp.299-308

2 - Sur Sicore en particulier

RIVIERE, P. (1994). Le système de codification automatique SICORE.
Working Paper, *Conférence des Statisticiens Européens, Séminaire ISIS 94,* Bratislava, Mai 1994

RIVIERE, P., SICORE, système général de chiffrement automatique, *Actes des Journées de Méthodologie statistique, 18-19 octobre 1995, INSEE-Méthodes* n°59-60-61, pp.143-185

RIVIERE P., SICORE, un outil et une méthode pour le chiffrement automatique à l'INSEE, *Courrier des Statistiques*, n°74, août 1995

SCHUHL P., SICORE, the INSEE Automatic Coding System, *Proceedings de l'ARC, Annual Research Conference*, Census Bureau, Washington, mars 1996

DESCHAMPS F., RIVIERE P., Codage automatique (1) : SICORE, *le Recensement de la Population 1999, Préparation (I), INSEE-Méthodes* N°79-80, pp. 267-299

3 - Sur l'enquête Emploi du Temps

Enquête sur les emplois du temps 1985-1986 : " Du Discours spontané des enquêtés à l'élaboration d'une nomenclature d'activités " Document de travail - novembre 1990

Note à l'attention de M Maurin - P.Rivière - N°057/C530/PR/PR du 08/03/1995
Note à l'attention de M Maurin " Un codage automatique des libellés d'Emploi du temps " F Dumontier et J L Pan Ké Shon - N°160/F340 du 29/9/1995

Note à l'attention de M Maurin " Codage automatique des libellés d'emploi du temps au moyen de Sicore - 3ème test-résumé " - F Dumontier et J L Pan Ké Shon - N°2/F340 du 21/1/1997

Note à l'attention de D Guillemot " 4ème test Codage Sicore-EdT " - F Dumontier et J L Pan Ké Shon - N°17/F340 du 5/2/1998

4 - Sur l'enquête Budget de Famille

Présentation dans le cadre de la réunion du réseau de concepteurs d'enquêtes - CAE - UMS - 26 février 1999 - " Enquête Budget de Famille 2000 : la préparation de la codification par SICORE " - F. Deschamps et S. Destandau

Présentation dans le cadre de la réunion du réseau de concepteurs d'enquêtes - CAE - UMS - 25 juin 1999 - " La qualité dans la future enquête Budget de Famille" - N. Cérani, S. Destandau et H. Fréchou

Les enquêtes sur le budget des ménages dans l'Union européenne - Méthodologie et recommandations pour l'harmonisation - 1997 - Eurostat -

Annexe N°1 : Sicore et ses connaissances...

Sicore est un logiciel de codification automatique créé par et pour l'INSEE dans les années 1990.

L'objet de ce papier n'est pas de présenter Sicore dans sa globalité, mais de comparer la manière dont il a été utilisé lors de la préparation de l'enquête Emploi du Temps 1998-1999 et lors de celle de Budget de Famille 2000-2001[35].

D'ailleurs, pour plus d'approfondissement sur la logique et l'architecture interne de Sicore, le lecteur pourra se référer aux articles cités en bibliographie.

1. Le codage automatique : de quoi s'agit-il ?

1.1 Coder

Assez simplement, le codage est l'action de transformer un texte en un code. Dans le contexte qui nous intéresse ici, il s'agit donc de convertir un produit de dépense consigné par l'enquêté vers le code de la nomenclature correspondant.

En théorie, tous les libellés devraient être codés. En pratique, cela n'est pas toujours le cas, et ceci pour plusieurs raisons. Tout d'abord, le libellé peut ne pas correspondre à ce à quoi on s'attend.

Par exemple, " CADRE " ne suffit pas à cerner une PCS et " ALPES " ne permet pas d'attribuer un code département.

Ensuite, le libellé peut ne pas contenir les informations suffisantes.

Dans le Val-de-Marne, deux communes commencent par " VILLENEUVE " : " VILLENEUVE-SAINT-GEORGES " et " VILLENEUVE LE ROI ". Ainsi, le seul libellé " VILLENEUVE " ne nous permettra pas de coder, même si nous savons que la commune en question est dans le Val-de-Marne.

Enfin, la nomenclature peut comporter des flous que chacun résoudra à sa manière, même si le résultat final du codage est non homogène. Nous verrons des exemples de ce cas plus tard.

En un mot, coder n'est pas facile, surtout lorsque l'on sait que **tous** les libellés doivent avoir un code à la fin.

[35] Pour cette enquête, nous nous intéresserons quasi uniquement à la variable " produit de la dépense " ; mais d'autres variables y sont codées : d'une part la CS contenue dans le tronc commun et d'autre part la variable " lieu de la dépense " qui, plus simple que la variable " produit de la dépense " ne nous intéressera pas dans ce papier.

1.2 Coder automatiquement

Chiffrer automatiquement, c'est faire coder le libellé par une machine. En tout état de cause, ce choix de la codification automatique de variables au sein d'une chaîne d'enquête n'est pas neutre.

En effet, ce choix induit des changements majeurs dans le calendrier de l'enquête, dans la présentation des supports de ces variables et dans son déroulement. Il faut d'abord prévoir en amont du codage automatique une saisie de qualité aussi bonne que possible, ce qui est relativement coûteux. Ensuite, comme le codage automatique n'est jamais efficace à 100%, il convient de mettre en place un dispositif permettant de reprendre les libellés non codés par la machine à la main (grâce à ce que l'on appelle un " outil de reprise " élaboré spécialement) ; ici encore cette opération est coûteuse manuellement et informatiquement, même si le volume de ces libellés est moindre que si l'ensemble des libellés était à coder manuellement.

Evidemment, les difficultés de codage que nous avons évoquées plus haut se retrouvent ici, et de manière plus forte encore : si la personne qui paramètre la codification automatique a du mal à arbitrer des codes, la machine ne pourra pas faire mieux.

2. Sicore : comment fonctionne-t-il ?

2.1 Principe général

Le principe de codage de Sicore est, en théorie, assez simple : en entrée, il reçoit les libellés de la saisie et en sortie, il renvoie le résultat de son codage (qu'il y ait un code ou pas).

Pour cela, Sicore agit en deux temps : d'abord, il normalise le libellé de la saisie de manière à le simplifier et l'homogénéiser à ce qu'il connaît (ou croit connaître). Ensuite, il tente de reconnaître ce libellé simplifié ; il s'agit là de l'étape de codage à proprement parler.

Toutes ces références dont on parle ici sont ce que l'on appelle en langage Sicore les *bases de connaissances*. De ce fait, il y a une totale séparation entre le programme et les connaissances relatives à une variable : les programmes sont toujours les mêmes quelle que soit la variable à coder. Mais cela implique évidemment que lorsque les connaissances d'une variable n'existent pas, il faut les écrire depuis le début.

2.2 La phase de normalisation

La première étape est dite de *normalisation*. Elle comporte plusieurs actions successives qui utilisent une partie des bases de connaissances de chaque variable. Ces actions successives sont :

↳L'élimination des caractères blancs, c'est à dire des caractères qui seront remplacés par un espace.

Exemples : l'apostrophe, la virgule, le point d'interrogation, ...

↳ L'élimination des caractères vides, c'est à dire des caractères qui seront éliminés

Pratiquement, seul le point est déclaré pour traiter les abréviations

↳ L'élimination des mots vides, c'est à dire des mots qui seront éliminés. A la différence de ce qui précède, cette rubrique est bien spécifique à la variable traitée.

Exemples pour " produit de dépense " : " L ", " PROMOTION ", " LOT ", ...

↳ L'utilisation des synonymes, c'est à dire des mots (ou groupes de mots) qui seront remplacés par d'autres mots (ou groupes de mots). Cette action a particulièrement un rôle d'entonnoir, dans le sens où elle permet, en plus de son action simple de synonymisation *(" AGNEAU " = " MOUTON " par exemple)*, de traiter des éventuelles fautes d'orthographe et/ou de saisie *(" AGNAU " = " AGNEAU ")*. Il convient alors d'être très attentif dans l'ordre dans lequel les synonymes sont rentrés dans Sicore.

↳Le calibrage, c'est à dire donner au libellé obtenu à l'issue des étapes précédentes une longueur fixée, en limitant le nombre de mots et la taille des mots *(exemple : 5 mots de 10 caractères)*. Lorsque les mots sont trop courts ou trop peu nombreux, on complète par des blancs.

2.3 Le fichier d'apprentissage

Lors de l'étape de codage, nous avons dit que Sicore essayait de rapprocher le libellé à coder de quelque chose qu'il reconnaît ou croit reconnaître. En disant cela, nous avons fait directement référence au *fichier d'apprentissage* qui est sans doute la partie des bases de connaissances la plus importante et la plus intuitive. Il s'agit d'un fichier contenant sur chaque ligne un libellé et le code associé à ce libellé.

Dans un premier temps, en fait avant la normalisation, le fichier d'apprentissage est dit *brut* : c'est celui sur lequel les améliorations sont effectuées, pour des soucis de meilleure lisibilité.

Dans le cas des pays et nationalités, le Fab a cette allure :

```
99101 208 208 COPENHAGUE
99101 208 208 ARHUS
99101 208 208 DANEMARK
99101 208 208 DANOIS
99438 212 212 DOMINIQUAIS
99438 212 212 DOMINIQUE
99438 212 212 ROSEAU
99408 214 214 REPDOMINICAI
99408 214 214 REPU DOMINICAINE
99408 214 214 REPUBDOMINIC
99408 214 214 REPUBLIQUE DOMINICAINE
99408 214 214 SAINT     DOMINGUE
```

Le code est sur l'ensemble des 13 premiers caractères, le libellé correspondant suit. On se rend compte ici que plusieurs libellés correspondent au même code ; en revanche, il n'est pas envisageable qu'un même libellé soit associé à plusieurs codes différents, même si cela peut exister dans les faits. Par exemple, une personne se déclarant de nationalité " BASQUE " peut tout à fait être dans les faits, espagnole ou française.

Enfin, le Fab peut comporter des *mots joker*. Il s'agit de mots dont la totalité des caractères peuvent être égaux à n'importe quel caractère (lettre ou chiffre). Cette technique, certes risquée, permet de traiter, par exemple, tous les libellés commençant par un certain groupe de mots. Par exemple, tous les libellés de profession commençant par " DIRECTEUR D'ECOLE " peuvent être traités, même si l'enquêté a rajouté le nom de l'école dans laquelle il est directeur (" DIRECTEUR DE L'ECOLE VICTOR HUGO ") ou la commune[36] (" DIRECTEUR DE L'ECOLE COMMUNALE DE MONNETIER ").

En pratique, et spécifiquement sur les variables qui nous intéressent ici, le Fab est la partie des bases de connaissances la plus travaillée et celle qui s'enrichit le plus au cours du temps.

2.4 L'utilisation éventuelle d'information supplémentaire

Or, il arrive que dans certains cas, pour certaines variables, on ne puisse coder directement un libellé. Sans même parler du cas de la PCS, on ne peut pas toujours

[36] Ceci n'est pas tout à fait exact : les directeurs d'école primaire ou maternelle, publique ou privée sont en 4214, à l'exception des directeurs d'école à classe unique classés en 4211. Toutefois, cette subtilité de déclaration est difficilement exigible de l'enquêteur et de l'enquêté.

coder une commune à partir du libellé si on n'a pas le département dans lequel est située la commune[37]. Dans de tels cas, on utilise de l'information supplémentaire sous forme de *variables annexes* que l'on traite dans des *règles logiques*. Le fichier des règles logiques est une nouvelle partie des bases de connaissances.

Dans le cas de la commune, on regarde toujours le département déclaré. Pour la PCS, 14 variables peuvent être nécessaires au codage. Dans les faits, on ne prend en compte que celles nécessaires au regard du libellé.

2.5 Les paramètres d'apprentissage : la mise en œuvre de la qualité du codage

A ce stade, on pourrait être tenté de croire que toutes les connaissances nécessaires au codage de la variable sont décrites. Ceci n'est pas tout à fait exact. Pour pouvoir coder le plus rapidement et avec la meilleure qualité possible, on définit ce que l'on appelle des *paramètres d'apprentissage*. Lorsque Sicore va apprendre toutes les connaissances écrites jusqu'ici, il va découper les libellés du fichier d'apprentissage, après normalisation, en groupe de deux lettres : les *bigrammes*.

De manière à vérifier la qualité du code attribué par Sicore, on définit des bigrammes de redondance, c'est-à-dire des bigrammes que l'on souhaite vérifier quoi qu'il arrive. Une fois le code attribué, Sicore va scrupuleusement vérifier tous ces bigrammes de redondance entre le libellé à coder et le libellé qu'il a de plus proche dans les connaissances qu'il a apprises.

Il suffit alors qu'un seul bigramme du libellé à coder soit différent du même bigramme du libellé le plus proche dans les connaissances apprises pour que l'on décrète qu'il y a erreur de redondance. Dans ce cas, même si Sicore renvoie un code, on considère le libellé comme non codé. En pratique, les 2 ou 3 premiers bigrammes des (2 ou 3) premiers mots sont souvent pris comme bigrammes de redondance ; cela dépend du fichier des connaissances de chacune des variables.

Prenons un exemple concret sur la variable " produit de la dépense ". Initialement, les trois premiers bigrammes des deux premiers mots du libellé normalisé sont redondants. Cela donne, pour 4 mots de 12 lettres (ou 6 bigrammes) chacun :

Soit maintenant le libellé brut suivant " une bouteille d'assouplissant " qui fait partie du Fab. Normalisé, ce libellé devient " BOUTEILLE ASSOUPLISSANT ". Imaginons maintenant que Sicore doive coder le libellé brut suivant " une

[37] A titre d'illustration, il faut savoir qu'il existe trois " MARSEILLE " en France

bouteille de soupline ". Normalisé, ce libellé devient " BOUTEILLE SOUPLINE ". Pour un codeur manuel, ce libellé ne posera aucun problème. Pour Sicore, il ne pourra être codé car, si le premier mot " BOUTEILLE " ne posera pas de problème, Sicore ne reconnaîtra pas le deuxième : dans le contexte décrit précédemment, ce libellé ne passera pas le contrôle de redondance[38].

De plus, d'autres paramètres d'apprentissage permettent à l'utilisateur d'agir sur la rapidité de codage et sur d'autres points de la phase de codage. Ces paramètres ne seront pas développés ici car ils ne nous intéressent pas particulièrement. De plus, l'ensemble du paramétrage n'est pas immédiat à mettre au point, et se fait par tâtonnement.

Tous ces paramètres d'apprentissage permettent à l'utilisateur de jouer sur un des grands paradoxes du codage automatique : le dilemme efficacité – fiabilité ; autrement dit, plus on code, moins on code juste. Il est facile de coder à 100% : il suffit de renvoyer l'ensemble des libellés sur un seul et unique code, via les mots joker. Arbitrer ce dilemme est beaucoup plus délicat et dépend très généralement des moyens alloués, notamment en reprise des libellés non codés.

2.6 La phase d'apprentissage

Une fois que toutes ces connaissances sont écrites et consignées, nous pouvons les faire apprendre, au sens littéral du terme, par Sicore lors de l'*apprentissage*. Lors de cette phase, Sicore prend en entrée les libellés normalisés du fichier d'apprentissage, les découpe en bigrammes et construit un *arbre de questionnement* selon les paramètres d'apprentissage définis par l'utilisateur. Nous ne nous attarderons pas ici sur la construction de cet arbre, même si, en théorie, beaucoup de choses sont à expliquer. Le lecteur espérant plus d'explications se reportera aux ouvrages cités en bibliographie.

[38] En pratique, les choses sont plus complexes du fait de la multiplicité des lignes dans le fab, y compris celles commençant par le mot " bouteille " : jus d'orange, eau, alcool, ...

2.7 Le codage à proprement parler

Une fois ces connaissances apprises, le codage à proprement parler peut commencer. Sicore prend alors un des libellés qu'on lui donne à coder, le normalise, le découpe en bigrammes et parcourt ses connaissances jusqu'à éventuellement trouver un libellé qui s'en rapproche. Trois cas généraux peuvent se présenter :

① S'il n'en trouve pas, on parle d'échec de codage, Sicore passe au libellé suivant.

② Si Sicore trouve un code, il effectue les contrôles de redondance que l'utilisateur lui a donné l'ordre de faire. Dès qu'un contrôle échoue, Sicore indique qu'il y a eu un problème de redondance, et donne le code qui aurait été trouvé s'il n'y avait pas eu de problème[39].

③ Si tous les contrôles sont satisfaisants, Sicore renvoie le " bon " code, en tous les cas, un code qui est pris comme étant bon. Les études de la qualité de codage détermineront éventuellement si certains codes sont justes ou pas.
En cas d'utilisation de variables annexes, le principe est le même : Sicore, après avoir traité le libellé suivant le précédent schéma parcourt les règles logiques de Sicore pour traiter l'information supplémentaire remplie par l'enquêté.

3. Une organisation spécifique autour de Sicore est indispensable

Il est assez naturel d'entrapercevoir le fait que choisir d'insérer Sicore au sein d'un traitement d'enquête est loin d'être neutre. C'est la raison pour laquelle, dès le déroulement du projet Sicore, il est apparu la nécessité d'imposer des règles d'organisation en cas d'utilisation de Sicore.

Cet aspect organisationnel revêt essentiellement deux composantes[40] : la gestion (élaboration éventuelle et mises à jour) des bases de connaissances et la définition des acteurs autour de Sicore avec la distribution des rôles.

[39] Malgré tout, dans de tels cas, le libellé est considéré comme non codé.

[40] En fait, un troisième élément est très important : l'insertion de Sicore au sein de la chaîne de traitement de l'enquête ; mais cet aspect ne nous intéresse pas ici, nous ne le traiterons pas.

3.1 La gestion des bases de connaissances

Le premier aspect met en place un processus cyclique appelé *boucle Sicore* dans deux cas : lorsque la variable est utilisée de manière permanente (comme la PCS-CS) ou lorsque la période d'enquête est suffisamment longue (comme pour les deux enquêtes qui nous intéressent ici).

En effet, pour qu'une application de codage automatique demeure efficace à long terme, il faut qu'elle soit vivante : le langage évolue, de nouvelles expressions apparaissent, et les nomenclatures changent également (apparition de nouveaux métiers, fusion de communes, ...). Si ce n'est pas le cas, les gestionnaires de la reprise qui traitent les libellés non-codés automatiquement récupèrent à chaque fois les mêmes mots, les mêmes termes, ce qui entraîne une démotivation de leur part ; et un surcoût non négligeable pour l'enquête du fait que des libellés pourraient être traités automatiquement.

Ensuite, il faut s'assurer que les connaissances de la variable ne contiennent pas d'erreurs. En effet, malgré toute l'attention portée au travail d'élaboration des connaissances et toutes les précautions que l'on peut prendre, des erreurs peuvent être mises en évidence au sein des bases de connaissances.
De plus, pour les deux variables qui nous intéressent ici (« activité quotidienne » et « produit de dépense »), nous verrons plus loin que les conditions de codage étaient initialement tout à fait similaires.

Le principe de cette boucle Sicore est assez simple : améliorer les bases de connaissances existantes pour coder plus et/ou mieux[41], en s'appuyant sur des résultats de codage précédents : on doit s'assurer que le codage précédent n'est pas trop mauvais, et faire en sorte que le nombre de non codés diminue progressivement. Toutes ces consignes assez naturelles à imaginer restent assez vagues. C'est là toute la difficulté de la mise à jour des bases de connaissances : il n'existe pas de règle miraculeuse qui améliorerait systématiquement le codage. On peut se contenter de fournir des pistes méthodologiques : tirer des échantillons de libellés codés et extrapoler la qualité résultante, essayer de traiter les principales fautes d'orthographe et/ou de saisie en rajoutant des synonymes, traiter en priorité les libellés non codés qui sont les plus fréquents après normalisation, ...

Cette boucle Sicore amène donc des changements des bases de connaissances. Ces changements sont possibles à plusieurs niveaux : synonymes, fichier d'apprentissage, paramètres d'apprentissage... Ils sont également possibles dans les deux sens, dans le sens d'une extension de ces bases, comme dans celui d'un rétrécissement : en effet, en cas de qualité jugée trop médiocre, on peut se rendre compte, par exemple, qu'un synonyme a un double sens non envisagé initialement et décider de l'éliminer.

[41] Cela dépend de l'option qui est prise sur l'arbitrage efficacité – fiabilité.

A la fin de cette phase d'amélioration, des procédures de vérification sont mises en place. Les plus couramment utilisées sont au nombre de deux : le codage du Fab sur lui-même qui doit arriver sur un codage total sans le moindre problème et le codage d'un précédent fichier d'enquête qui ne doit pas voir l'efficacité et la qualité changer du tout au tout par rapport au précédent codage.

3.2 La structure des acteurs gravitant autour de Sicore

La deuxième composante essentielle de l'aspect organisationnel lié à Sicore concerne les acteurs fondamentaux de la codification automatique. En théorie, ils sont au nombre de cinq.

Le système Sicore s'organise autour d'un *expert Sicore*. Celui-ci est le point d'entrée, le relais, pour tout statisticien qui voudrait utiliser le codage automatique dans son application. Il centralise les bases de connaissances, gère le réseau des experts de variables et les groupes de travail correspondants, prend en compte les demandes sur le codage automatique et conseille les utilisateurs, organise les formations, spécifie les évolutions de l'outil, gère la communication interne et externe. L'expert Sicore est en lien étroit avec l'*informaticien Sicore* qui est chargé de la maintenance évolutive de l'outil, de l'intégration de Sicore dans les chaînes de traitement d'enquête, et, parallèlement, de la documentation informatique de Sicore.

Viennent ensuite les utilisateurs de Sicore. D'abord, il y a l'*expert variable* (un expert pour chaque variable codée par Sicore) est la cheville ouvrière du système. C'est lui qui a la lourde tâche de construire et de mettre à jour la base de connaissances qui l'intéresse. L'*informaticien de l'enquête* doit écrire la chaîne de traitement informatique, et donc y insérer les traitements induits par l'utilisation de Sicore, ceci en coordination avec l'informaticien Sicore. Enfin, le statisticien d'enquête est la personne à l'origine de tout : c'est lui qui passe la commande du codage automatique et supervise toute son enquête, en particulier le codage automatique. C'est par exemple cette personne qui peut arbitrer entre efficacité et fiabilité.

4. Les différentes utilisations de Sicore antérieures à EdT et à BdF

L'une des particularités de Sicore est son aspect transversal : il est utilisable pour toutes les enquêtes ménages pour coder des variables diverses. Avant les deux enquêtes dont il est question ici, Sicore codait plusieurs variables. On comptait alors deux variables simples, sans variable annexe : le département, et le code pays/nationalité.

Ensuite, nous avons le cas un peu plus compliqué de la commune : il faut absolument que le département soit renseigné et on compte en plus une variable annexe : la date à laquelle on veut que le code se réfère. Si aucune date n'est renseignée, le codage donne le code connu le plus récent : de ce point de vue, la date est une variable annexe non nécessaire.

Enfin, la PCS représente la variable la plus lourde à gérer : 14 variables annexes, de natures très diverses et pour certaines déjà issues de codification précédente, en plus du libellé, si possible assorti d'une indication de grade pour les employés de la fonction publique

Toutes ces variables évoquées précédemment, avant le lancement de l'enquête EdT, avaient la particularité d'avoir déjà leur base de connaissances et étaient utilisées pour des enquêtes diverses : toutes celles utilisant le tronc commun des enquêtes ménages, l'Etat-civil, des enquêtes régionales de déplacement, ...

Nous allons voir qu'avec les deux enquêtes EdT et BdF, les choses sont un peu différentes : rien n'existait avant la phase de préparation des deux enquêtes

Annexe n°2 - L'enquête Emploi du Temps

Depuis 1964, l'INSEE réalise des enquêtes Emploi du Temps environ tous les dix ans. La dernière enquête de 1998-1999, comme les précédentes, comporte 8 vagues de 6 semaines chacune, réparties sur un an, du 15 février 1998 au 15 février 1999. 12 000 logements sont concernés.

L'objectif principal de ces enquêtes est de savoir quelles sont les occupations quotidiennes des individus des ménages ordinaires et leur durée.

Les supports de l'enquête de 1998-1999 sont tous papier :

- la fiche-adresse
- les questionnaires
- les carnets journaliers

Le carnet journalier auquel nous nous intéressons ici permet de décrire toutes les activités d'un même jour (24 heures) de chaque individu de 15 ans et plus du ménage interrogé.

- les semainiers

L'enquêteur se déplace deux fois chez le ménage. La première fois, il remplit le questionnaire ménage, explique le remplissage du carnet (en faisant un exercice avec le ménage), fixe le jour de remplissage du carnet, administre les questionnaires individus avec les personnes (de 15 ans et plus), présentes, et laisse au ménage les carnets à remplir et un " semainier " de travail pour les actifs occupés. Il revient une seconde fois chez le ménage, pour ramasser les carnets, les contrôler avec le ménage et pour administrer les questionnaires individus qui ne l'avaient pas été lors de la 1ère visite.

1. L'organisation du traitement des documents

Tout le travail de saisie, d'apurement, de lancement du codage automatique et de reprise de codage de tous les documents a été confié au GSAS de la DR des Pays de la Loire.
Chaque personne de cet atelier pouvait revenir au dossier papier complet. Des réunions fréquentes avaient lieu au sein de l'équipe pour mettre en commun les problèmes rencontrés qui étaient alors transmis, soit au responsable d'enquête, soit à l'expert variable, par le responsable.

Les dossiers papier au niveau ménage étaient groupés par lots.

2. L'outil de reprise

Une fois les carnets saisis et apurés, le lancement de la codification automatique était réalisé par une personne de l'équipe qui traitait alors les reprises. Pour ce faire, un outil a été mis au point par l'informaticien de l'enquête (*Isabelle Rebourg* du CNIA).

Lors des reprises effectuées par les personnes du GSAS grâce à cet outil, apparaissaient non seulement le libellé à coder, ses 9 variables annexes, mais aussi les intitulés des activités précédentes et suivantes, pour mieux situer l'activité dans la journée

Grâce à un menu spécifique de cet outil, l'expert variable avait la possibilité de voir l'avancement du codage et de demander des listes d'activités codées et non codées triées selon plusieurs critères.

L'ensemble de l'enquête comportait **429 678 libellés à coder**, la plupart étant des activités principales (316 097) et le reste des activités secondaires (environ 27% de l'ensemble des activités).

ANNEXE N°3 – Exemple de remplissage du carnet journalier de l'Enquête Emploi du Temps

Vous comprendrez facilement comment procéder en regardant les exemples et en appliquant les consignes suivantes :
- Notez vos occupations **régulièrement et plusieurs fois dans la journée** afin d'éviter les oublis.
- Notez vos occupations que vous délimiterez par une accolade aux heures de début et de fin de chaque occupation
-Il doit y avoir une utilisation de **toutes les heures de la journée**; on peut très bien réfléchir, se reposer, parler, ou attendre quelque chose pendant une partie de la journée, et il **faut le marquer**

	Marquez vos différentes occupations de la journée en indiquant les heures de début et de fin de chaque occupation à l'aide d'accolades dans la colonne de gauche		**Faites-vous autre chose en même temps ?** *(lecture, conversation, radio, TV...)*	**lieu ou trajet** une réponse possible					**en présence de qui** 3 réponses possibles				**votre activité est dans un but** une réponse possible			
				Chez soi	Lieu de travail	A l'extérieur	Trajet domicile travail	Autres trajets	Seul	Autre personne du ménage	Amis, voisin, parenté collègue	Autre personne	Personnel ou pour son ménage	Professionnel	Pour un autre ménage	Associatif
7h00		je dors		0	1	2	3	4	1	2	3	4	1	2	3	4
10				0	1	2	3	4	1	2	3	4	1	2	3	4
20				0	1	2	3	4	1	2	3	4	1	2	3	4
30				0	1	2	3	4	1	2	3	4	1	2	3	4
40		je fais ma toilette	radio	0	1	2	3	4	1	2	3	4	1	2	3	4
50		je m'habille	radio	0	1	2	3	4	1	2	3	4	1	2	3	4
8h00		je prépare le petit déjeuner	conversation	0	1	2	3	4	1	2	3	4	1	2	3	4
10		je déjeune	TV	0	1	2	3	4	1	2	3	4	1	2	3	4
20				0	1	2	3	4	1	2	3	4	1	2	3	4
30		je fais la vaisselle	TV	0	1	2	3	4	1	2	3	4	1	2	3	4
40		je range la cuisine	TV	0	1	2	3	4	1	2	3	4	1	2	3	4
50		je fais le ménage		0	1	2	3	4	1	2	3	4	1	2	3	4
9h00		je vais au travail avec ma	conversation	0	1	2	3	4	1	2	3	4	1	2	3	4
10		femme et un voisin		0	1	2	3	4	1	2	3	4	1	2	3	4
20				0	1	2	3	4	1	2	3	4	1	2	3	4

ANNEXE n°4 - Résumé de la codification automatique des activités journalières de l'enquête Emploi du Temps

Date	Lignes à coder	Environnement	Efficacité		Qualité
		Sicore	activité pincipale	activité secondaire	activité principale
Construction et test du Fab					
mars 1995	construction FAB recherche codes multiples codage FAB sur FAB *Test 1* 19 échantillons de 200 lignes test de codage	199 codes d'activité FAB=3918 lignes Pas de règles Pas de synonymes 6 big sur les 2 1er mots 19 FAB complément	100% codés **82% codés simples** 18% multiples (1/5) **40.4% codés (en moy)** 32.4% codés simple 8.0% codés multiples		76.1%
été 95	*Test2 1er essai* 2500 lignes issues 84 carnets saisis de enquête 1986 *Test 2 2ème essai*	199 codes d'activité FAB= 8500 libellés **650 synonymes** 6 big sur 1er et 2ème mot **50 règles** + 2 big sur 3ème mot	**67% codés** dont 13% multiples **70% codés** dont 3% multiples		81% 80%

Tests réels sur le terrain					
juin 96	test3 3253 libellés issus test 2 DR variables annexes	100 codes d'activité FAB = 9500 libellés 1000 synonymes **introduction de jokers**	**80% codés**		90%
juin 97	test4 4270 libellés issus test 2 DR	amélioration FAB= 10500 lignes 1650 synonymes	**83%**		93,5%
sept 97	test5 4220 libellés issus de deux DR	1er codage même environnement que test4	**78;5%**		80,5%
		2ème codage environnement amélioré	**81.6%**	98% estimée	96.1%
			activités principales **et** secondaires **86%**		
Enquête réelle 15/2/98 au 15/2/99					
du 15/2/99 au 29/398	vague 1 55 410 libellés à coder	FAB=11000 libellés 2000 synonymes 77 règles nomenclature 103 postes	**90.4%**		
avril 98	recodification 8 vagues 429 678 libellés à coder (dont 27% secondaires)	FAB=13400 libellés 2388 synonymes 99 règles nomenclature 139 postes	**92.2%** 89.9% principales	98.2% secondaires	

Annexe n°5 : L'enquête Budget de Famille

L'enquête Budget de Famille ressemble étrangement à sa petite sœur l'enquête Emploi du Temps. En effet, BdF est une enquête ancienne (depuis 1965 sous ce nom) et réalisée tous les 5 ans par l'INSEE. Elle comporte 8 vagues de 6 semaines réparties sur un an. La dernière en date s'étale de mai 2000 à mai 2001 pour la version métropolitaine. La taille de son échantillon est supérieure à celle d'EdT : 20 000 logements métropolitains.

Son objectif principal consiste à mesurer le plus précisément possible les dépenses, les consommations et les ressources des ménages français.

Pour cela, elle repose sur 2 types de support :

- des questionnaires posés au ménage sous CAPI
- des carnets de dépenses papier.

Avant de réaliser l'enquête en grandeur nature, deux tests ont eu lieu sur le terrain grâce à des enquêteurs de 5 Directions Régionales différentes en novembre 1998 et juin 1999.

1. L'organisation du traitement des carnets papier

Compte tenu du nombre de FA de l'enquête BdF supérieur à celui de l'enquête EdT(20 000 contre 12 000), du nombre de dépenses par ménage (par rapport au nombre d'activités dans un carnet journalier), un choix d'organisation bien différent de celle adoptée pour l'enquête EdT en ce qui concerne le traitement des carnets a été nécessaire.

1.1 La saisie des carnets de dépense

Sur l'ensemble des vagues de l'enquête, le volume à traiter a été estimé ainsi :

Nombre de FA de l'échantillon initial[42] métropole + DOM	23 000
Nombre de ménages répondants métropole + DOM	16.000 (estimation par rapport à l'enquête de 94-95)
Nombre moyen estimé de dépenses par ménage	134
Nombre estimé de dépenses à saisir	**2 144 000 dépenses**

[42] Initialement nous n'avions pas prévu d'échantillon de réserve métropolitain. Cet échantillon comporte 2000 FA en plus des 18 000 FA standards pour la métropole

Compte tenu de ce volume, tous les ateliers de saisie en exercice en mai 2000 ont dû saisir les dépenses des carnets. Etant au nombre de 13 contre 18 DR de collecte, des regroupements de DR ont été effectués selon la taille des GSAS.

Par exemple, le GSAS de Toulouse a pris en charge la saisie des carnets de la DR du Languedoc Roussillon et de ceux de l'Auvergne en plus de ceux de leur DR.

1.2 La reprise des carnets de dépenses

Pour la reprise des carnets de dépenses, le volume à traiter et l'organisation de la saisie a eu des conséquences non négligeables sur l'organisation de la reprise qui est en fait elle aussi bien différente de celle choisie pour EdT.

** une estimation de volume à reprendre démoralisante*

Taux estimé de reprise des produits en 2000 (4 vagues) Nombre estimé de produits à reprendre en 2000	40% 857 600 produits à reprendre
Taux estimé de reprise des produits en 2001 (4 vagues) Nombre estimé de produits à reprendre	20% 428 800 produits à reprendre
Taux estimé de reprise des magasins Nombre estimé de magasins à reprendre	10% 112 000 magasins à reprendre

** l'organisation de la saisie , une contrainte pour la reprise*

Compte tenu de la multitude des sites de saisie, 2 options ont été prises :

- d'une part, nous sommes parties du principe qu'il fallait déplacer le moins possible les carnets après leur saisie de manière à ce que les personnes effectuant la reprise puissent les consulter.
Les 13 DR effectuant la saisie ont donc été chargées aussi de la reprise, l'organisation au sein de la DR étant laissée entièrement libre. En réalité, pour la majorité des 13 DR (8), la reprise des carnets s'est effectuée au sein des DEM. Dans 4 DR, le travail de codage manuel est partagé entre la DEM et le GSAS et enfin dans seulement 1 DR (Nantes, là où les carnets d'EdT ont été entièrement traités) toute la reprise a été confiée au GSAS.

- d'autre part, la codification automatique n'est pas lancée par les différentes équipes assurant derrière la reprise des rejets comme pour EdT mais par l'informaticien de l'enquête au niveau central (CNI de Lille). Il assure ensuite la mise à disposition des dépenses non codées automatiquement auprès de ces 17 équipes ((GSAS+DEM)*4 + 8 + 1) dans les 13 DR de reprise afin qu'elles les codent manuellement grâce à un outil informatique spécifique à BdF.

2. L'outil de reprise

En plus de constituer le Fab produits, tout le travail de constitution des connaissances a permis de choisir une présentation des carnets, mais aussi de spécifier l'outil de reprise qui est utilisé par les 17 équipes en DR durant plus d'un an, pour coder manuellement des dépenses non codées automatiquement.

- En effet, d'une part, il nous a paru intéressant de conserver l'information du type de codage (codé, redondant, échec) c'est à dire l'écho de Sicore dans l'affichage des dépenses. Ainsi face à un produit (ou un magasin) redondant, le codeur peut s'aider du code proposé par Sicore pour coder manuellement.
- D'autre part, l'outil n'affiche pas seulement le produit rejeté (c'est-à-dire redondant ou en échec) mais bien l'ensemble des informations relatives à ce produit c'est à dire
  - s'il provient du tableau de dépenses :
    * la quantité et unité quand elle existe,
    * le montant de la dépense,
    * le type de magasin.
  - s'il provient d'un ticket de caisse :
    * le montant de la dépense,
    * le type de magasin.

Bien entendu, toutes les autres informations relatives à l'identification du ménage, de l'individu et de la date de l'enquête peuvent apparaître à la demande du codeur afin qu'il évite de revenir au carnet papier.

Le codeur peut donc être confronté au cas où le produit est à coder mais le magasin est codé automatiquement. Dans ce cas, le libellé du magasin peut l'aider à coder le produit.
Exemple : " 1 - COUPE - 220 FRANCS - COIFFEUR "

ANNEXE n°6 – Exemple de remplissage du carnet de dépenses de l'Enquête Budget de Famille

Nième jour de collecte

Jour de la semaine : *Mercredi*

Date : *18-11-1998*

TICKETS DE CAISSE

	Intermarché XXXXXX		
GINI 10x20 CL 6 1/1.95	18.70		
VOLVIC	11.70		
OUTIL, AUTO MAG	169.00	→	→ *housse siège voiture*
JOUET LOISIRS	25.80		
MIEL ACACIA	15.30		
COQUIGRAIN 65 /70x6	6.40	→	→ *oeufs*
MAJESTY BOEUF	2.10	→	→ *alimentation chat conserve*
BANGA 4x20 CL	6.95		
FEMME	49.90	→	→ *slip femme*
RICORE 250G			
BATONNETS COTON	3.45		
GENIE GEL	11.80		
BTE PYRENE	21.95		
			→ *cadeau*

Nième jour de collecte

Jour de la semaine : *Mercredi*

Date : *18-11-1998*

TABLEAU DES DÉPENSES

Quantité Unité	Nature de la dépense	Montant en francs	lieu de dépense
1	*facture EDF*	*782.20*	*prélèvement*
2 billets	*train*	*1250*	*SNCF*
1	*magazine*	*10.50*	*presse*
1	*pull over cadeau*	*250*	*Monoprix*
3 places	*cinéma*	*150*	*Gaumont*

Annexe n° 7 : Glossaire

BdF : enquête Budget de Famille

CAPI : Collecte Assistée par Informatique

CNE : Centre National d'Exploitation dans une DR

CNI : Centre National Informatique
(Aix - Lille - Orléans - Nantes - Paris)

COICOP-HBS : nom de la nomenclature européenne (Eurostat) des enquêtes Budget de Famille

DEM : Division Enquête Ménage (SES) dans une DR

DG : Direction Générale

DR : Direction Régionale

EdT : enquête Emploi du Temps

FA : Fiche-Adresse

Fab : Fichier d'apprentissage brut

GSAS : Atelier de saisie (SAR) dans une DR

PCS : Catégorie socio-professionnelle à 4 chiffres

CS : Catégorie socio-professionnelle à 2 chiffres