4.2 - La codification automatique : MCA et SICORE

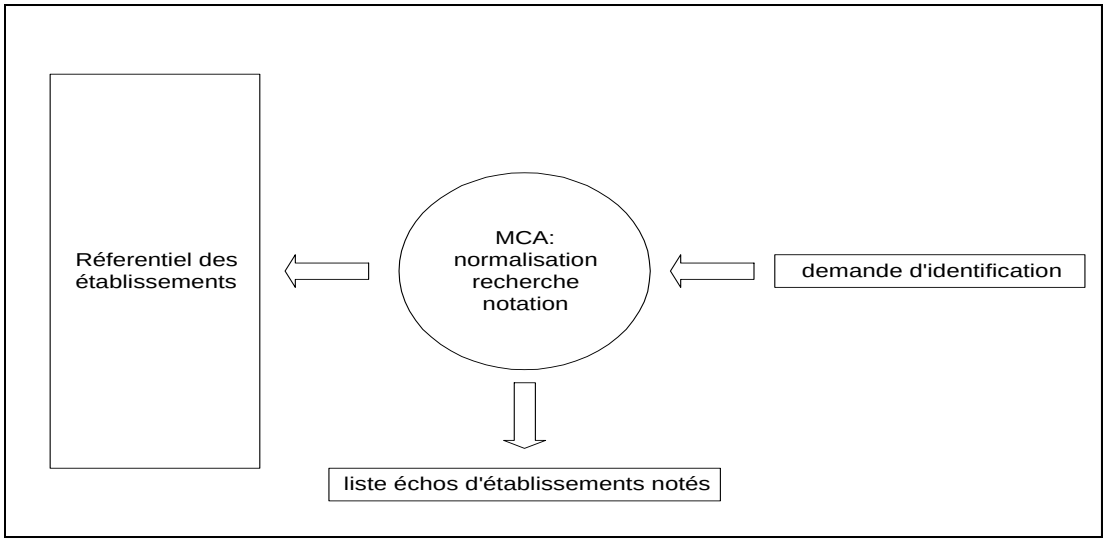
La codification automatique utilise deux logiciels différents selon la variable à codifier ; la Mise en Concordance Automatique (MCA) pour le chiffrement de l'activité de l'établissement employeur, et SICORE pour le chiffrement des autres variables (pays, nationalité, commune, profession, activité antérieure), et comme aide pour le chiffrement de l'activité actuelle. L'activité servant au chiffrement de la profession, ce chiffrement est réalisé en premier.

1 - La MCA :

Le protocole adopté pour le chiffrement de l'activité économique des bulletins du recensement exploite une procédure d'identification développée au CNI de Nantes : la Mise en concordance automatique (MCA). Cette dernière est utilisée dans un premier temps **en batch** afin d'assurer la codification automatique de masse des bulletins puis **en mode interactif** (dans RECAP) afin de traiter les bulletins non codés automatiquement.

Les étapes du processus d'identification sont :

- la constitution d'un référentiel d'établissements
- la normalisation d'une demande d'identification
- la recherche d'échos dans le référentiel d'établissements
- la notation des échos trouvés dans le référentiel



Chaque demande d'identification (correspondant à une déclaration dans un bulletin) doit dans un premier temps pouvoir être normalisée afin de préparer la comparaison avec le référentiel des établissements. Par la suite, le module de recherche, permet d'extraire du référentiel les échos vraisemblables. Enfin, la notation attribue à ces derniers une note en vue de la décision d'identification.

1.1 - Le référentiel des établissements

L'identification concerne la recherche d'établissements. Le référentiel est constitué de :

- tous les établissements actifs à la date de référence de la collecte (le 15 janvier de l'année N)

- la liste des établissements des pays frontaliers (Allemagne, Belgique, Luxembourg, Suisse et Monaco)

La géographie du référentiel correspond au dernier millésime du COG (Code officiel géographique).

1.2 - La normalisation

La normalisation permet d'éliminer dans une demande d'identification les mots ou caractères considérés **vides** dans le Nom et Raison Sociale et l'adresse (virgules, « de », « des », « le », « au », parenthèses…).

La normalisation permet également de faire converger des mots sémantiquement équivalents à l'aide de synonymes. Par exemple, les mots « Mairie », « Hôtel de ville », « Municipalité », « Ville » ou « commune » sont équivalents et doivent faire l'objet d'une **synonymisation** identique.

De même, la normalisation gère les mots en pluriel (équivalent à leur singulier). Par exemple, « Ford industrie » est équivalent à « Ford industries ».

1.3 - La recherche

Deux modes de recherche élémentaires sont utilisés pour le codage :

- **recherche sur la base de la déclaration du nom raison sociale**
- **recherche sur la base de la déclaration de l'adresse de localisation**

La recherche sur la base de la déclaration du NRS permet de retrouver tous les établissements de NRS, enseigne, sigle ou noms usuels approchant la déclaration.

La recherche de type adresse permet de retrouver tous les établissements localisés à une adresse précise.

1.4 - Les données Sirène à rapatrier

Pour chaque écho trouvé en recherche dans le référentiel, le logiciel récupère des variables sirene niveau établissement (numéro Siret, NRS, la tranche d'effectif salarié, Apet, nom usuel, adresse sirene, …) et quelques variables niveau entreprise (Catégorie juridique, Apen, Tranche d'effectif salarié, recme). Ces variables sont nécessaires pour un bon codage de la profession.

1.5 - La notation

La notation des échos contient deux éléments :

- une note de vraisemblance au niveau de l'adresse
- une note de vraisemblance au niveau du nom

1.6 - L'analyse des échos à conserver

En fonction des notes obtenues par chaque écho, une procédure propre au recensement décide automatiquement les échos à conserver. La décision est différente selon que l'on est en batch ou l'objectif est de trouver avec certitude le bon écho, et éviter ainsi de passer par la phase reprise, ou en mode interactif ou il s'agit de proposer à l'opérateur une liste d'échos, classés par ordre de vraisemblance, afin qu'il retrouve facilement le bon écho parmi ceux-ci.

1.7 - L'utilisation de SICORE activité pour valider les échos douteux

Afin d'améliorer les taux de chiffrements automatiques de l'activité en batch, lorsque la note du premier écho proposé par la MCA n'est pas suffisante pour décider de le conserver, un chiffrement du libellé d'activité déclaré est réalisé à l'aide du logiciel Sicore. Si le code activité chiffré est identique au code activité fourni par le premier écho de la MCA, alors l'écho est conservé et le chiffrement est validé. Dans le cas contraire, le questionnaire est renvoyé en reprise.

1.8 - Taux de chiffrement automatique

Le taux de chiffrement automatique de l'activité de l'établissement employeur est compris entre 40 et 45%. Les raisons du non codage proviennent essentiellement de fautes d'orthographe, d'une divergence entre le nom et raison sociale de Sirène et celui qui a été déclaré, une mauvaise localisation de la commune du lieu de travail, ou un problème d'adressage.

2 - SICORE :

Sicore est un système de codification automatique de libellés qui utilise un fichier de référence appelé fichier d'apprentissage (ou base de connaissance) lui servant d'exemples de chiffrement. Le libellé à chiffrer est ainsi comparé aux libellés contenus dans le fichier d'apprentissage et, lorsque le libellé est reconnu, le code associé est retenu. Le code proposé peut être un code intermédiaire regroupant plusieurs réponses possibles. Le bon choix est alors réalisé en utilisant des variables annexes et des règles logiques. Par exemple, pour le chiffrement de la profession, la connaissance de la fonction exercée et de l'activité de l'établissement dans lequel travaille la personne, permet dans de nombreux cas d'aboutir au bon code de la nomenclature.

2.1 – Le fichier d'apprentissage

Il contient les libellés de référence pour le chiffrement qui serviront de base de comparaison avec le libellé à chiffrer. Chaque libellé de référence est associé à un code de la nomenclature. Il y a un fichier d'apprentissage spécifique pour chaque variable à coder. Ordres de grandeur du contenu des différents fichier d'apprentissage :

- Pays et nationalités : 1 800 libellés de référence ;
- Communes : 49 100 libellés de référence ;
- Activités : 20 000 libellés de référence ;
- Profession : 27 000 libellés de référence ;

2.2 - Les règles de normalisation

Les libellés sont normalisés afin de faciliter la comparaison. Les libellés de référence sont normalisés lors de la constitution du fichier d'apprentissage, alors que les libellés à coder sont normalisés avant d'effectuer la comparaison. Les grandes étapes de la normalisation sont :

- Enlever ce qui ne donne pas d'information (caractères et mots vides : virgules, tirets, …).
- Résumer l'information (donner des expressions synonymes : Saint=St, …).
- Calibrer l'information (nombre et longueur des mots pour la comparaison).

Exemples de règles de normalisation :

- Chiffrement des Communes : 60 mots vides, 40 expressions synonymes, 5 mots maxi répartis en 1 mot de 2 caractères maxi (N° de département), 1 mots de 14 caractères maxi et 3 mots de 12 caractères maxi.

- Chiffrement de la profession : 300 mots vides, 780 expressions synonymes, 6 mots maxi répartis en 1 mot de 2 caractères maxi et 5 mots de 12 caractères maxi.

2.3 - La comparaison des libellés

La comparaison des libellés se fait par bigrammes (groupes de deux lettres). SICORE compare les bigrammes des libellés à chiffrer aux bigrammes des libellés de référence. Cette comparaison se fait sur des bigrammes prioritaires (1 à 8) puis sur les autres bigrammes si nécessaire. Les bigrammes prioritaires sont ceux qui sont les plus significatifs pour la reconnaissance d'une expression (en général les 3 premiers bigrammes des 3 premiers mots en commençant par le deuxième bigramme de chaque mot).

1 - Premier exemple de libellé à coder après normalisation :

50	**CO**	**UT**	**AN**	**CE**	S
1	**3**	**2**	**4**	**5**	

Le fichier d'apprentissage comprend:

03 COUTANSOUZE	03089
50 COUTANCES	50147

Réussite du codage dès le 5ème bigramme.

2 - Deuxième exemple de libellé à coder après normalisation :

78	**MA**	**NT**	**ES**
1	**3**	**2**	**4**

Le fichier d'apprentissage comprend:

78 MANTES JOLIE	78361
78 MANTES GASSICOURT	78361
78 MANTES VILLE	78362

Echec du codage : plusieurs résultats possibles.

2.4 - L'utilisation des variables annexes

La comparaison des libellés ne suffit pas toujours pour coder, certains codes proposés par le fichier d'apprentissage étant des codes intermédiaires ou précodes. A partir de ce précode, une table contenant les règles logiques entre plusieurs variables appelées variables annexes, permet d'aboutir à des codes différents en fonction des valeurs prises par ces variables.

Les variables annexes utilisées pour le chiffrement de la profession sont : la position professionnelle, le département du lieu de travail, la fonction principale exercée, l'activité principale de l'établissement, le nombre de salariés pour les chefs d'entreprise, l'orientation des productions agricoles, la variable PUB (distinction public/privé), le statut de l'emploi, la taille de l'entreprise et le sexe de la personne en emploi.

2.5 – La validation du codage

En plus du chiffrement réalisé, SICORE fournit également un code retour qui décrit la façon dont s'est déroulé le processus de codification. En fonction du code retour fournit, la codification est validée ou non.

Les codes retour Sicore et leur signification :

Code retour	Signification	Signification détaillée
CCS	Réussite	Réussite sans utilisation des variables annexes
RCS	Réussite	Réussite avec utilisation des variables annexes
RC*	Réussite sous réserve de qualité	Réussite mais une (ou plusieurs) variable annexe nécessaire était à valeur manquante
CCM	Réussite mais en codage multiple	Plusieurs codes possibles ont été trouvés
C_R	Échec partiel (erreur de redondance)	Éventuel codage pouvant être considéré comme douteux
C_C	Échec	Non reconnaissance du libellé
R_V	Échec	Problème de transcodage avec une des variables annexes
R_R	Échec	Cas non prévu dans les règles logiques
R_B	Échec	Boucle dans le parcours des règles logiques

2.6 - Les taux de chiffrement automatiques

Les taux de chiffrement automatique sont de :

99% pour les pays et la nationalité
97% pour les communes
88% pour la profession antérieures
82% pour la profession actuelle.