La reprise manuelle des échecs de codification automatique de la profession en PCS 2020

I – Dans quels cas effectue-t-on de la reprise manuelle de codification ?

Le protocole de codification de la profession en PCS 2020 s'appuie sur une matrice de codification. Elle indique comment sont codés en PCS 2020 environ 5 800 libellés de profession (dont près de 5 400 sont proposés dans la liste de collecte). Dans cette matrice, les libellés de profession sont présentés en ligne, les variables annexes en colonne. La combinaison de ces informations détermine les codes de profession, au croisement des lignes et des colonnes. Néanmoins, cette matrice n'est pas exhaustive de toutes les professions ni de toutes les combinaisons de variables annexes. C'est pourquoi, il est nécessaire d'effectuer de la reprise manuelle de codification dans certains cas :

- Quelques croisements de la matrice ne proposent pas de code PCS lorsque la combinaison du libellé et des modalités des variables annexes ne permet pas d'attribuer de code, sans information contextuelle (comme le grade pour les libellés « militaire » ou « policier » par exemple). Dans ce cas, le croisement contient la lettre 'r'.

- D'autres croisements fournissent un code PCS au niveau regroupé des professions (3 premières positions du code PCS). Par exemple, lorsque le libellé correspond à un emploi du secteur public, et que la fonction publique d'appartenance (État, Territoriale, Hospitalière) est nécessaire pour coder au niveau le plus fin, le code PCS se terminera par un 0. Ces cas pourront être codés sur 4 positions en reprise manuelle si des informations contextuelles (comme les coordonnées de l'établissement employeur) permettent de les coder plus finement. Sinon, le résultat de la codification automatique aboutissant à un codage partiel sera conservé.

- Enfin, le protocole de collecte du libellé prévoit la possibilité de saisir son libellé de profession en clair s'il n'est pas trouvé dans la liste. Dans la plupart des cas (aux synonymes près après normalisation du libellé), ces libellés devront être traités par un gestionnaire de reprise.

Les libellés collectés hors liste ou plus généralement les libellés non prévus dans la matrice de codification ou renvoyant le code 'r' ou un code se terminant par '0' devront être traités manuellement[1]. Un pôle spécialisé de la Direction Régionale de l'Insee Bourgogne-Franche-Comté localisé à Besançon, réalise les opérations de reprise pour les enquêtes ménages de l'Insee. Selon les besoins, il traite également des enquêtes des services statistiques des ministères en offre de service.

II – Les grands principes de la reprise manuelle

Lorsqu'un libellé de profession n'est pas codé automatiquement, il faut le traiter manuellement. Pour cela, plusieurs grands principes doivent être respectés et la maîtrise de la nomenclature PCS est requise pour ce travail.

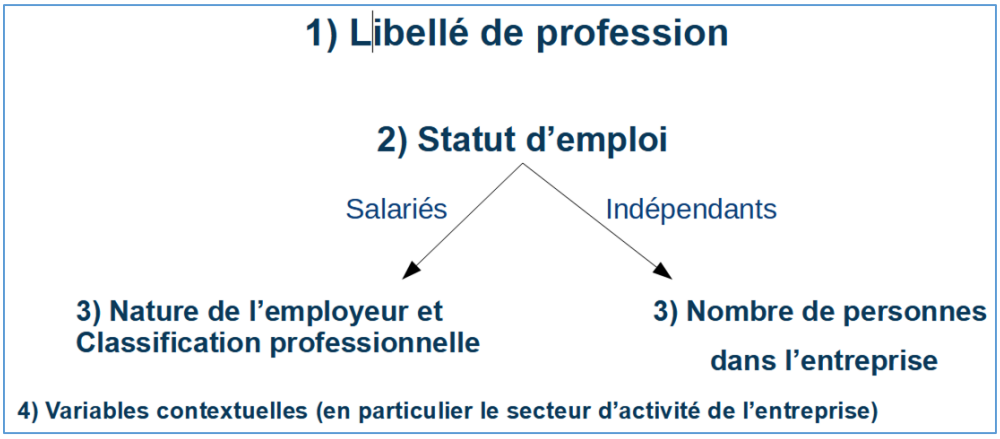
II – 1 - Respect des grands clivages de la nomenclature PCS 2020.

La connaissance et le respect des grands clivages de la nomenclature PCS 2020 aboutissent à une utilisation hiérarchisée des variables annexes ou contextuelles à disposition. Toutes les variables (libellé, statut, nature de l'employeur, classification professionnelle, taille de l'entreprise) n'ont en effet pas la même importance pour le codage de la profession. Les informations apportées par chacune de ces variables sont ainsi hiérarchisées :

[1] Ou par méthode d'imputation probabiliste ou par combinaison de méthode d'imputation et de reprise manuelle selon la qualité de codage attendue. À l'heure actuelle, aucun modèle standard n'a été développé par l'Insee.

1. **Le libellé déclaré prime sur toutes les variables annexes**. Cela se traduit par le fait que dans un certain nombre de cas, le codage du libellé ne respectera pas les informations apportées par les variables annexes. Par exemple : le libellé « Président Directeur Général (PDG) d'entreprise de l'industrie » sera toujours codé dans la CS 23, même si l'enquêté déclare un statut de salarié ou une taille d'entreprise inférieure à 10 personnes.
2. **En deuxième position par ordre de priorité, le statut d'emploi de l'individu l'emporte sur les autres variables annexes** en cas d'incohérence entre elles. Cette hiérarchie traduit le clivage transversal de la nomenclature PCS entre indépendants et salariés.
3. **Pour les salariés, il n'y a ensuite pas de hiérarchie entre la nature de l'employeur et la classification professionnelle** qui structurent toutes les deux la nomenclature au niveau des groupes socioprofessionnels. **Pour les indépendants, la taille de l'entreprise structure la nomenclature des groupes socioprofessionnels 1 et 2**.
4. **Les variables contextuelles interviennent en dernier lieu**. Ces variables, qui apportent de l'information complémentaire sur la situation professionnelle de l'enquêté, ne sont pas requises pour la codification en PCS 2020 mais peuvent être utilisées en reprise lorsqu'elles sont disponibles. Notamment, la connaissance de l'activité économique de l'employeur ou la fonction d'un salarié permettent dans de nombreux cas de coder au niveau le plus fin de la nomenclature.

Le schéma suivant résume la hiérarchie des variables de codification :



II – 2 - Connaissance des règles de construction de la matrice de codification

Afin d'harmoniser la reprise manuelle avec la codification automatique, il convient de s'appuyer sur les règles qui ont été édictées pour la construction de la matrice de codification.

Une des principales règles de codification est le codage par défaut à retenir pour les libellés pouvant être codés dans plusieurs groupes ou catégories socioprofessionnelles selon la valeur de leur statut d'emploi, de la nature de leur employeur ou de leur position professionnelle. **En PCS 2020, en l'absence d'information sur la classification professionnelle, le codage par défaut est le codage au plus probable[2] et s'effectue au mode**. Il n'y a ainsi pas de règle générale mais une règle spécifique pour chaque libellé (ou par groupe de libellés qui désignent le même métier ou un métier de la même famille) et non par rubrique entière qui mélange souvent plusieurs familles de métiers. Ces situations devraient toutefois être rares, la réduction du nombre et la simplification des variables nécessaires au codage de la nomenclature devant limiter les réponses imprécises ou manquantes.

Cette règle est appliquée pour chacune des 4 variables annexes pour déterminer le codage à appliquer par défaut d'information sur l'une ou plusieurs de ces variables. Pour déterminer ces règles de codage par défaut pour les quelque 5 800 libellés de la matrice de codification, les codeurs se sont appuyés sur le mode calculé sur les données déclarées dans l'enquête de recensement 2017. Ils ont pu ainsi déterminer **une variable dite « nature du libellé » qui détermine le codage en l'absence d'information sur le statut d'emploi de l'individu et la nature de l'employeur pour les salariés. Ils ont également pu définir un codage par défaut au plus probable pour la classification professionnelle, lorsque cette information était disponible.**

[2]Par rapport à la PCS 2003, cette règle a évolué puisque le codage par défaut se faisait dans la classification la plus basse possible ce qui entraînait une déformation systématique de la structure socioprofessionnelle vers le bas.

Par exemple, la règle du codage au plus probable se traduit par :

- en l'absence de la variable relative au statut de l'enquêté, le libellé « Parfumeur-créateur » sera codé en indépendant,
- en l'absence de la nature de l'employeur, le libellé « chercheur en biologie » sera codé en salarié du public plutôt qu'en salarié du privé,
- en l'absence de la position professionnelle, le « plombier » salarié sera codé en ouvrier qualifié plutôt qu'en ouvrier peu qualifié.

Par ailleurs, afin de garder une cohérence transversale de codage entre des libellés de profession portant une information sur la responsabilité ou la compétence, des règles ont été édictées sur le niveau de codage possible de ces libellés. De plus, la matrice de codification a été construite en distinguant les libellés de profession selon des filières d'emploi. Au sein de chacune des filières, des règles spécifiques ont été édictées afin d'assurer une cohérence interne. Cette matrice ainsi qu'un document synthétique expliquant le principe de ces règles avec des exemples sont disponibles sur le site dédié à la nomenclature (nomenclature-pcs.fr dans la rubrique Coder).

Afin de respecter l'ensemble de ces règles, notamment celle du codage au mode en l'absence partielle ou totale des variables annexes, un rapprochement des libellés à reprendre avec ceux présents dans la matrice de codification est nécessaire.

Dans de nombreux cas, ce rapprochement est cependant insuffisant. Leur résolution nécessite alors une expertise du cas et/ou l'application de consignes de reprise. Pour traiter les enquêtes ménages de l'Insee mais également en offre de service d'autres enquêtes sous maîtrise d'ouvrage des services statistiques des ministères, ce sont des gestionnaires spécialisés du pôle PCS qui effectuent cette codification manuelle.

Pour effectuer ce travail, les gestionnaires disposent d'une application spécifique qui leur permet de traiter l'ensemble des variables d'emploi (activité et profession). Dans cette application, toutes les informations (libellé, description des tâches si le libellé est flou, variables annexes et contextuelles collectées) utiles pour un codage précis peuvent être consultées.

Dans un souci d'harmonisation de leur pratique, les gestionnaires sont formés au respect d'un ensemble de consignes spécifiques établies par le pôle PCS. Une documentation interne, mis à jour régulièrement, recense l'ensemble de ces consignes ainsi que des décisions prises au fil de l'eau ce qui permet de conserver une cohérence des pratiques de résolution des cas au cours du temps.