Direction générale
Direction de la méthodologie et de la coordination statistique et internationale
Unité « SSP Lab »

Dossier Suivi par :
LEROY Theo
Tél : 0187695533
Mèl : theo.leroy@insee.fr
MALHERBE Lucas
Tél : 0187695578
Mèl : lucas.malherbe@insee.fr
SEIMANDI Tom
Tél : 0187695516
Mèl : tom.saimandi@insee.fr

Note à l'attention de

Christel Colin (DSDS)
Sylvie Lagarde (DMCSI)

Montrouge, le 2 septembre 2021
N°2021_18494_DG75-L401

Objet : Fiche de lancement – Expérimentation SSP Lab – DMTR – DRTI

Codification automatique des professions dans la nomenclature PCS 2020 pour les futurs libellés en dehors de la liste

Sponsors :

Gwennaël Solard, Chef de la division Méthodes et Traitements des Recensements, DSDS, Amandine Schreiber, Cheffe de la Division Recueil et Traitements de l'information, DMCSI

Sujet :

La rénovation de la nomenclature des PCS en 2020 s'accompagne de la promotion d'un outil d'autocomplétion des libellés de profession dans une liste de libellés enrichis permettant le codage direct dans la case de la nomenclature ou des regroupements *ad hoc* complémentaires de la nomenclature. Le passage par cette liste de libellés enrichis rend caduque l'environnement Sicore de codage en PCS dès lors que l'outil d'autocomplétion est utilisé. Il n'est pas prévu de développer un moteur de règles complet similaire à Sicore PCS 2003. Or, l'outil d'autocomplétion ne sera disponible que pour des collectes informatisées et par ailleurs il comprend la possibilité de répondre « hors liste ». Pour pouvoir intégrer la nouvelle PCS 2020 dans le recensement de population à la collecte 2024 et être en mesure de coder aussi les bulletins papier tout comme les réponses informatisées « hors liste », un algorithme de codification automatique en PCS 2020 de ces bulletins doit être créé. L'objectif étant de maintenir le taux de codifications correctes, tout comme un taux d'envoi en reprise manuelle à des niveaux similaires à l'existant. Des travaux préalables du DMS et du SSP Lab ont confirmé l'intérêt d'une approche par apprentissage supervisé (fasttext). Suite au report de l'enquête de recensement de 2021, des gestionnaires ont annoté en PCS 2020 avec double codage et arbitrage, 119 000 bulletins issus de l'EAR 2020, en mobilisant une interface de labellisation mise en place par DRTI pour l'occasion. Cet échantillon annoté en PCS 2020 servira d'échantillons d'entraînement/validation pour entraîner/tester les algorithmes d'apprentissage supervisé.

Objectif :

Le but de l'expérimentation est alors de tester, comparer différents modèles d'apprentissage statistique et méthodes de prétraitement pour le codage des professions dans la nomenclature PCS 2020 à partir du libellé de profession ainsi que des variables annexes utilisées lors de la phase d'annotation (statut de l'employeur, position professionnelle, etc.) et d'en retenir le plus performant. Un souci particulier portera sur la constitution de l'échantillon d'entraînement, notamment en tenant compte de la qualité des données annotées disponibles. Il est aussi envisagé d'être capable de recourir à des codages à 2-3 ou 4 positions selon la qualité / précision de l'information disponible. L'évaluation des différentes méthodes retenues sera fondée sur des critères de performance (efficacité et fiabilité) mais devra tenir compte des contraintes opérationnelles pour que ces méthodes soient intégrées si les résultats sont concluants dans le système d'information de l'Insee.

Mode de travail :

Les partenaires consacreront en moyenne une journée par semaine à l'expérimentation. Point d'avancement tous les mois avec les superviseurs.

Groupe impliqué dans l'expérimentation :

Superviseurs : Gwennaël Solard (DMTR), Amandine Schreiber (DRTI), Elise Coudin (SSP Lab)

Experts métiers : Maxime Exavier (DMTR), Souheil Benmebkout (DMTR)

Experts méthodologiques : Stéphanie Himpens (SSP Lab), Lucas Malherbe (SSP Lab) et Théo Leroy (DRTI)

Référent informatique et sécurité pour l'utilisation du datalab: Yves-Laurent Benichou

Rôles

Les superviseurs définissent les objectifs de l'expérimentation, assistent aux démonstrations des avancées, jugent de la conformité des travaux à leurs attentes et redéfinissent les propriétés au besoin.

L'équipe de réalisation (expert métiers et experts méthodologiques) définissent et réalisent conjointement les étapes et les livrables dans le calendrier prévu ci-dessous.

Le référent informatique et sécurité pour le datalab accompagne les expérimentateurs concernant la demande et l'utilisation de la plateforme datalab. Il est responsable du suivi de la sécurité concernant le projet sur la plateforme datalab s'assure de la mise en œuvre effective des mesures de sécurité du ressort de l'équipe projet, assure un reporting auprès de la DSMR et, le cas échéant, tient à la disposition d'une mission d'audit les éléments de preuve sur la conformité de l'expérimentation eu égard aux engagements pris.

Ressources

Les données annotées seront disponibles dans un espace sécurisé sous AUS V3 et, diminuées des informations qui pourraient être indirectement identifiantes (à savoir la raison sociale), sur le datalab. La mise à disposition des données sur le datalab ne pourra se faire que si l'expérimentation a reçu un avis d'homologation sécurité favorable.

Le SSP Lab met à disposition le lieu pour accueillir les expérimentateurs lors de sessions de travail en commun. Sous réserve d'avis d'homologation sécurité positif, la DSI (Unissi)

met à disposition les accès au datalab et proposera son assistance pour le paramétrage des ressources correspondantes.

Calendrier prévu et types de rendus

Modèle de codification automatique des bulletins individuels du RP en PCS 2020 (T4 2021)

Bilan des estimations, tests et comparaison des modèles, amenant à choisir un modèle ainsi qu'une première base d'entraînement, recommandations adressées à la MOA concernant la façon d'implémenter et de mettre à jour le modèle et l'échantillon d'entraînement (T4 2021)

Présentation dans un séminaire dédié aux expérimentations liées au déplacement du RP (T4 2021 – T1 2022)

Présentation aux JMS (T1 2022) et document de travail méthodologique

La cheffe du département de la démographie

Signé : Valérie Roux

La cheffe de l'unité « SSP Lab »

Signé : Elise Coudin

Le chef du département de la méthodologie statistique

Signé : Patrick Sillard

Pour information :
Agents des départements SSP Lab, DMTR, et DRTI