Journées de méthodologie statistique de l'Insee 2022

APPARIEMENTS SÉCURISÉS DE DONNÉES PERSONNELLES : UN EXEMPLE POUR LA RECHERCHE

Yacine EL BOUHAIRI, Kamel GADOUCHE, Rémy MARQUIER (*)

(*) Centre d'accès sécurisé aux données

service@casd.eu

Mots-clés : appariements, NIR, hash, CSNS, réglementation

Domaine concerné : intégration des données, institutionnel

Résumé

Nous proposons une présentation sur les dispositifs d'appariements de données personnelles pour la recherche au travers de l'exemple FORCE associant les données de la DARES et de Pôle Emploi. Ces appariements, déterministes, se font *via* deux méthodes, fonction des identifiants utilisés : NIR (autrement appelé *numéro de sécurité sociale*) ou tryptique nom/prénom/date de naissance. L'articulation avec le dispositif en cours de construction à l'INSEE de NIR haché pour le cas particulier du service statistique public fera par ailleurs l'objet d'un focus.

Depuis quelques années, en particulier depuis la Loi pour une République numérique de 2016 (LPRN), le législateur encourage la mise à disposition de données administratives, y compris des données très détaillées (au niveau des individus ou des entreprises), issues tant des administrations centrales (DGFiP, SIES, DREES…) que des organismes investis d'une mission de service public (caisses de sécurité sociale notamment) à des fins de recherche scientifique ou de statistique.

Parallèlement, la même Loi pour une République Numérique a exprimé la volonté de faciliter les appariements de données personnelles sur la base du NIR pour les chercheurs. L'utilisation des données administratives et les appariements permettent d'une part d'alléger la charge de réponse des organismes ou individus interrogés lors d'enquêtes ou de remontées d'information formelles (ce qui rejoint le principe du « dites-le nous une fois »), d'améliorer la précision des indicateurs pour les enquêtes statistiques, et par ailleurs d'enrichir considérablement les analyses statistiques et de recherche.

Il n'en reste pas moins que depuis sa création, la CNIL est particulièrement vigilante à tout projet d'appariement sur de larges échelles, et spécialement ceux qui utilisent le NIR, imaginant *via* la LPRN un dispositif de « hachage » du NIR, destiné à transformer celui-ci en un code statistique non signifiant (CSNS) de manière irréversible. De premiers projets utilisant le CSNS ont été mis en œuvre, ce dispositif est désormais encouragé par l'ASP dans son délibéré du 22 septembre 2021. Les cas de la recherche scientifique et celui du service statistique public, sont distincts, le second pouvant disposer d'un NIR haché commun à toutes ses opérations.

Le dispositif permettant les appariements *via* le NIR haché a été formalisé par le décret n°2016-1930, repris ensuite par la loi Informatique et Libertés modifiée en 2018. Le schéma imaginé en lien avec la CNIL introduit deux tiers de confiance : le tiers chargé de réaliser le ou les algorithmes de hachage du NIR, et celui chargé de réaliser les appariements (cf. schéma joint). Le CASD joue le rôle de ce deuxième tiers de confiance dans le cadre du dispositif PIC-FORCE (DARES, Pôle Emploi) et dans le futur dispositif MIDAS (DARES, Pôle Emploi, CNAF).

Dans le cadre du dispositif FORCE, outre les appariements sur le NIR (fichiers FH et MMO), d'autres données ne disposent pas du NIR (c'est notamment le cas des bases I-MILO et BREST). Reconstituer le NIR avant appariement est délicat, cette reconstruction demandant un dispositif lourd et des informations identifiantes complètes : en plus du nom, du prénom et de la date de naissance, la zone géographique de naissance est requise. Dès lors que ces informations sont incomplètes et que le NIR n'est pas reconstitué, un appariement partiel est toujours possible en utilisant directement les informations identifiantes à disposition, *via* une méthode basée sur la distance de Jaro-Winkler ou Levenstein, et utilisant des seuils de distance après mise en forme homogène des identifiants. D'autres appariements du même type sont par ailleurs mis en œuvre par ajout de données administratives locales, fonction des projets de recherche. Des identifiants anonymes sont ensuite conservés pour retirer toute possibilité d'identification directe. Le travail débouche sur plusieurs tables de passage entre les différents identifiants, ainsi qu'à la construction d'un identifiant unique global.