Journées de méthodologie statistique de l'Insee 2022

APPARIEMENTS SÉCURISÉ DE DONNÉES PERSONNELLES : UN EXEMPLE POUR LA RECHERCHE

Yacine El Bouhairi, Kamel Gadouche, Rémy Marquier(*)

(*) Centre d'accès sécurisé aux données – CASD

yacine.elbouhairi@casd.eu kamel.gadouche@casd.eu remy.marquier@casd.eu

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

Domaines concernés : Histoire, Appariements, fusion de sources (*record linkage*), couplage « exact » de fichiers

Résumé

Depuis sa création en 1978, la CNIL est particulièrement vigilante à tout projet d'appariement sur de larges échelles, et spécialement ceux qui utilisent le NIR. La loi pour une République numérique (2016) propose un dispositif de « hachage » du NIR, destiné à transformer celui-ci en un code statistique (pour le service statistique public) ou spécifique (pour la recherche) non signifiant de manière irréversible. Outre les aspects historiques entourant les appariements sur la base du NIR, 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 ou tryptique nom/prénom/date de naissance. Pour compléter le panorama autour des codes non signifiants, 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.

Abstract

Since its creation in 1978, the CNIL (control authority for personnal data) is deeply concerned by projects of linkage on a large scale, especially those which use the NIR (French register number). Since Digital law in 2016 (loi pour une République numérique), it is possible to use an encrypted NIR for linkage, available for the national statistics administration (CSNS) as well as for researchers (CRNS). The operation is irreversible. Furthermore to historical point of views on linkage upon the NIR, we propose a presentation ok linkage operation for research through the example of the FORCE apparatus, which links DARES data (Ministry of employment) and Pôle Emploi data (unemployment administration). These determinists methods works in two ways : NIR or usage of other identifiers name/first name/date of birth. The apparatus for statistics administrations is briefly described in order to fullfill the panorama.

Introduction

Les appariements de microdonnées (de personnes ou d'entreprises) dans un but d'amélioration de la connaissance constituent un enjeu crucial pour la recherche scientifique et la statistique publique. En effet, ceux-ci permettent un enrichissement considérable des analyses et recherches, en liant des personnes sur des domaines où les données sont construites le plus souvent de façon étanche[1]. Il peut s'agir par exemple de relier les caractéristiques d'emploi avec la consommation de soins des personnes pour étudier les risques psycho-sociaux, ou encore de lier les bases de bénéficiaires de minima sociaux, d'employés et de demandeurs d'emploi pour évaluer les politiques publiques sur l'emploi. Outre cet enrichissement, les coûts associés aux appariements sont significativement moindres que ceux qui se révéleraient en conduisant des opérations de collecte multidomaines indépendantes, fondamentalement utiles par ailleurs. Le coût de ces opérations est d'autant plus important qu'il s'agit également d'assurer une représentativité statistique suffisamment robuste. Là encore, les appariements permettent de rester à une échelle importante pour un coût peu élevé. Ainsi, s'agissant des bases que l'administration détient, les populations recouvertes sont très importantes, la plupart du temps exhaustives sur leur champ, permettant, outre l'enrichissement des analyses proprement dite, une précision importante des résultats obtenus : estimateurs potentiellement sans biais et de variance nulle ou quasi-nulle, un atout considérable pour les statisticiens. Enfin, une conséquence heureuse des appariements : la charge de réponse des individus ou entreprises s'en trouve allégée (on respecte ainsi le principe du « dites-le nous une fois »).

A l'heure de la puissance décuplée des moyens informatiques et du big data, qui fait l'objet de beaucoup de confusions et nourrit conséquemment un nombre incalculable de fantasmes[2] , des appariements sur des bases de données gigantesques apparaissent techniquement et financièrement

[1] On parle le plus souvent d'architecture « en silos », qu'il faut « décloisonner ».

[2] Pour les exemples les plus caricaturaux, regardez n'importe quelle série policière grand public de ces cinq dernières années. Cela peut prêter à sourire mais beaucoup y croient, y compris à un haut niveau.

faisables, même si ce n'est pas forcément facile d'un point de vue juridique, suscitant une demande croissante.

Nous faisons le point dans cet article sur les possibilités d'appariements permises par la législation et son évolution récente, puis nous montrerons à travers l'exemple du dispositif FORCE mené par la DARES et Pôle Emploi en partenariat avec le CASD, le processus d'appariement de personnes physiques dans le cadre de projets de recherche.

1. Une législation et un cadre réglementaire qui ne cessent d'évoluer

1.1. Des appariements qui nécessitent d'avoir conscience des enjeux de protection de la vie privée

Le milieu des années 70 a été le théâtre d'un scandale retentissant autour du projet SAFARI (acronyme du Système Automatisé pour les Fichiers Administratifs et le Répertoire des Individus) consistant à chaîner les informations individuelles de personnes dans des domaines variés, sur la base du NIR[3]. A l'époque, l'affaire avait fait la Une du journal Le Monde sous le titre « SAFARI ou la chasse aux français », la peur du « fichage » de la population prenant le pas sur toute autre considération. S'en est suivi une période d'âpres discussions parlementaires, qui a donné naissance à la loi de 1978 relative à l'informatique, aux fichiers et aux libertés (loi dite « Informatique et libertés »). En résumé, la loi institue le fait qu'un organisme est en droit de traiter des données personnelles, mais pose un certain nombre de garde-fous pour éviter les dérives, en particulier la surveillance de masse. On parle alors de « respect de la vie privée ».

Le RGPD (adopté en 2016 au niveau européen et d'application française depuis le 25 mai 2018) a actualisé en quelque sorte la loi informatique et libertés, tout en renversant la logique des responsabilités : désormais et sauf cas particulier, il n'est plus nécessaire de déclarer un traitement à la CNIL pour le mettre en œuvre, il est par contre nécessaire de documenter sa conformité en cas de contrôle, en particulier en menant une analyse de risque pour les traitements considérés comme sensibles.

Le RGPD et la loi française établissent une classification des données personnelles. Ainsi, certaines données sont considérées comme très sensibles : il s'agit de données relatives aux droits humains (opinions philosophiques, religieuses, politiques, appartenance à un syndicat…) ou touchant directement à l'intimité de la personne (orientation et comportement sexuels, santé, caractéristiques génétiques ou biométriques…). Ces données doivent être manipulées sous certaines conditions quant à la finalité du traitement et la sécurité qui l'entoure. Les traitements (dont la collecte de données) associés à ces données sont interdits sauf exceptions limitativement énumérées.

Le NIR, dont on vient de rappeler la sensibilité, n'est pas cité en tant que tel dans le RGPD comme étant une donnée sensible, le règlement renvoyant aux éventuels textes nationaux des pays membres (art. 87 du RGPD). La législation française considère toujours cette variable comme hautement sensible et y dédie un article spécifique dans la loi informatique et libertés (art. 30) : en effet il eût été étonnant que cette variable ne fasse pas l'objet d'attentions particulières étant donné le contexte précédant la naissance de la loi informatique et libertés. Le NIR permet justement les appariements de façon très précise : ce numéro est unique et porté par la personne tout au long de sa vie ; il permet d'identifier de façon sûre un individu, au contraire du simple couple nom-prénoms ou, plus récemment, de la reconnaissance faciale.

Jusqu'à très récemment (2019 !), il n'était possible d'utiliser le NIR « en clair », c'est-à-dire sans transformation, pour la recherche et l'évaluation qu'en passant par un décret en Conseil d'Etat, procédure demandant d'une part un accès privilégié à l'administration d'Etat (cabinets ministériels), d'autre part force patience et un argumentaire extrêmement solide. De ce fait, si les appariements

[3] Ou numéro de sécurité sociale, pour les profanes.

sur la base du NIR étaient en théorie possibles techniquement, la réglementation en vigueur imposait des conditions tellement draconiennes que ce dispositif restait relativement rare et essentiellement mis en œuvre par les services statistiques de l'Etat.

La loi Informatique et Liberté modifiée en 2019 comporte un paragraphe spécifique aux utilisations particulières du NIR, en renvoyant vers un décret en conseil d'Etat la liste des responsables de traitement et les finalités pour lesquelles l'utilisation de cet identifiant est possible. La recherche scientifique n'est pas citée dans ce décret et même la statistique publique se trouve très contrainte pour l'utilisation du NIR en clair.

1.2. Une avancée permettant de contourner le NIR sans atténuer la qualité : les Codes non signifiant

Les possibilités d'utilisation du NIR pour la réalisation d'appariements dans la statistique publique ou pour la recherche en sciences humaines sont quasiment hors d'atteinte actuellement, sauf pour les opérations qui existaient préalablement au décret de 2019 (EDP, DADS, échantillons interrégimes…). Toutefois, d'autres procédures ont été permises dans les années récentes, facilitant les appariements via des processus dédiés qui, toujours en utilisant un numéro d'identification sûr, permettent de mettre en œuvre ces traitements sans devoir en passer par la case Conseil d'Etat.

Ainsi de la loi pour une République numérique (2016), comportant plusieurs dispositions visant à améliorer l'accès aux archives détenues par l'administration (dont les données), notamment pour les besoins de la recherche scientifique. Cette loi comprend un article dédié pour les appariements sur la base du NIR, en introduisant les concepts de « Code statistique non signifiant » et de « Code spécifique non signifiant » (art. 34 de la loi), applicables pour tout projet d'appariement sans passer par le NIR original. De premiers projets utilisant ces nouveaux codes ont été mis en œuvre, ce dispositif est désormais encouragé par l'ASP dans son délibéré du 22 septembre 2021.

Le décret n°2016-1930 détaille les dispositions applicables pour utiliser ces numéros identifiants, en distinguant deux cas : celui du service statique public (INSEE et Services Statistiques Ministériels – SSM) qui utilise le Code statistique non signifiant – CSNS - et celui de la recherche en sciences humaines qui utilise, lui, le Code spécifique non signifiant, qui sera appelé pour plus de lisibilité Code recherche non signifiant (CRNS[4]) dans le reste du document. Concrètement, le NIR est « haché » via un algorithme garantissant son unicité tout en empêchant tout retour en arrière. C'est ce NIR haché qui est utilisé pour la réalisation des appariements (Figure 1).

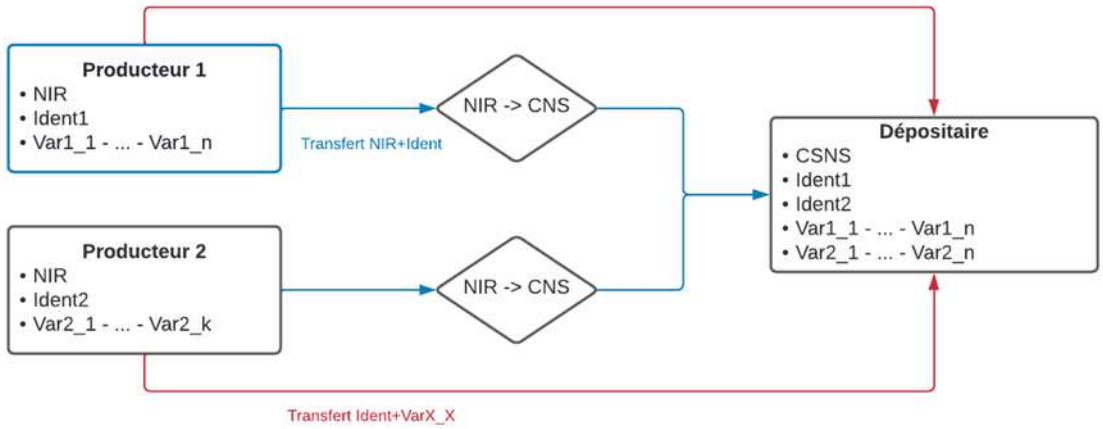

Figure 1 : processus de mise en œuvre du CNS pour un projet d'appariement

[4] Appellation d'origine non contrôlée, proposée par le CASD pour le langage courant.

Pour fonctionner, le mécanisme interdit toute construction d'une table de passage entre le NIR d'origine et le NIR haché.

Le NIR est chiffré via un algorithme lié à une clé dite de « salage », permettant de reproduire le résultat dès lors que la même clé est utilisée. Cette clé doit rester secrète et connue uniquement d'un tiers dit « de confiance ». La durée de vie d'un CNS est de dix ans, sauf en cas d'atteinte potentielle à la sécurité du processus, auquel cas le CNS doit être regénéré à l'aide d'une autre clé de salage immédiatement.

Dans le cas du SSP, le NIR haché est construit et mis à disposition par l'INSEE via un portail dédié[5]. Ce CSNS est valable pour l'ensemble du SSP et ne peut pas en sortir, il doit être conservé dans des environnements sécurisés (Figure 2). Il est important de noter que le NIR stocké dans les tables d'origine est détruit après validation du résultat de l'opération cryptographique par le service statistique ministériel qui en a fait la demande. De même, le CSNS est détruit par l'INSEE à chaque opération de constitution de celui-ci, après validation du résultat par le service statistique ministériel. Cette exigence avait été imposée par la CNIL pour éviter toute constitution de table de passage entre NIR et CSNS.

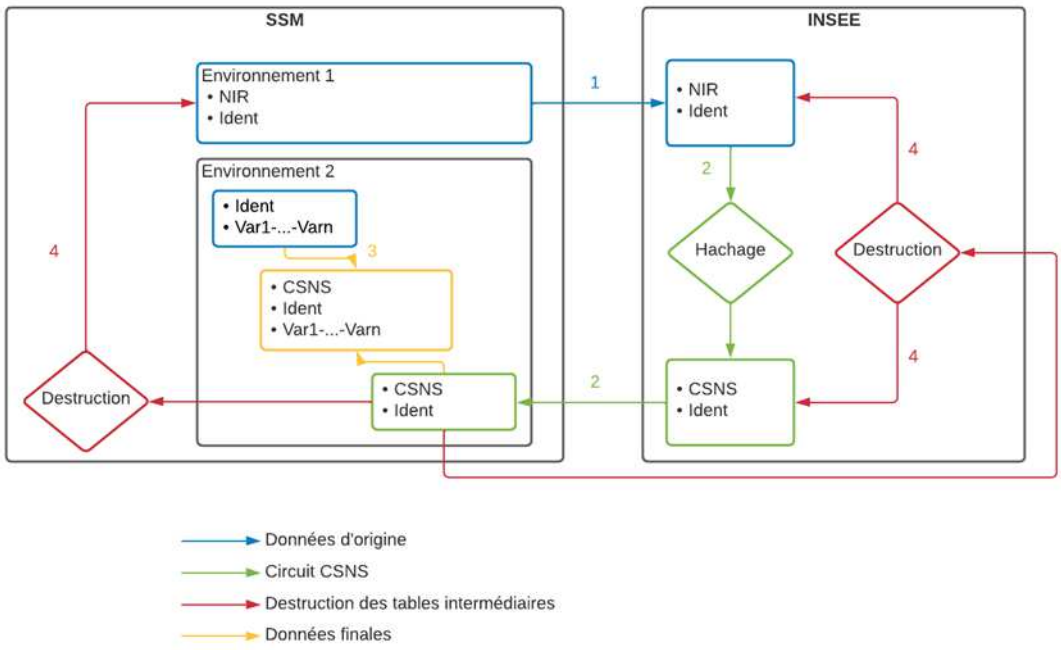

Figure 2 : cycle du CSNS dans le cas du SSP

Pour la recherche scientifique, non seulement le CRNS doit être généré par un premier tiers de confiance (qui n'est pas forcément l'INSEE cette fois-ci), mais un deuxième tiers de confiance entre en jeu pour réaliser les appariements demandés dans le projet de recherche. D'autre part, le CRNS permettant cet appariement est supprimé du fichier d'exploitation avant mis à disposition de la base appariée. Le CRNS dédié à un projet de recherche est lié uniquement à ce projet de recherche et ne peut pas servir à d'autres projets d'appariement, il y a nécessité d'utiliser un autre algorithme ou une autre clé de salage (Figure 3).

Pour le cas particulier du CASD, celui-ci peut mettre à disposition via son infrastructure le résultat de l'appariement.

[5] Ce portail vient d'être ouvert, une première demande de la DREES est en cours de traitement pour son enquête CARE-I.

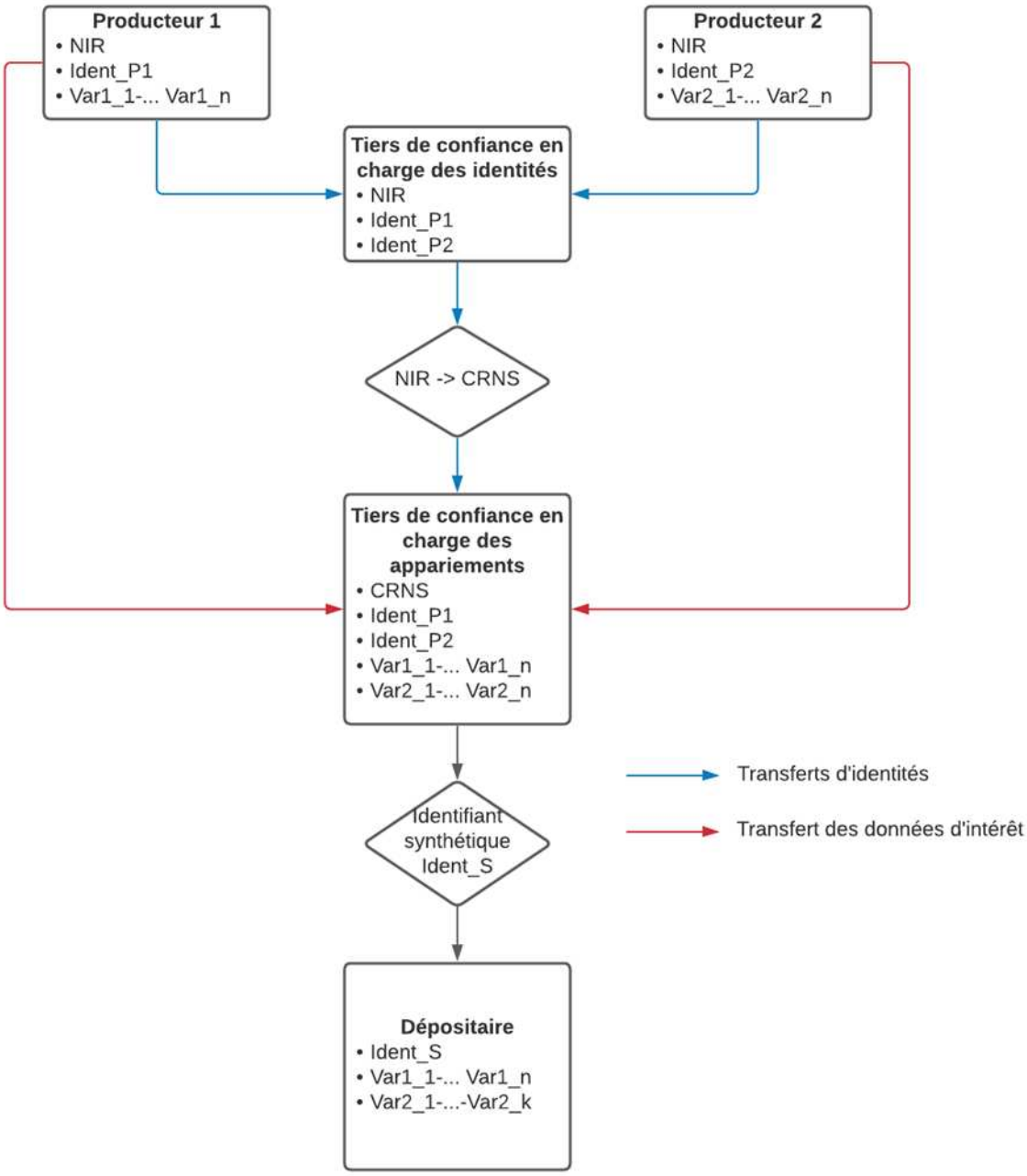

Figure 3 : projet d'appariement pour la recherche via le CRNS

2. Le projet FORCE : un appariement *via* le NIR haché mais pas seulement

Le dispositif FORCE (Formation, Chômage et Emploi) a été mis en place par la DARES et Pôle Emploi pour permettre l'évaluation du Plan d'investissement dans les compétences 2018-2022 (PIC). Il s'agit de permettre la reconstitution des trajectoires professionnelles de toutes les personnes ayant eu un contact avec le service public de l'emploi (missions locales, Pôle emploi) et/ou ayant suivi une formation professionnelle prise en charge totalement ou partiellement par les pouvoirs publics. Les appariements se font chaque trimestre et perdurent sur plusieurs années. Les données sont ensuite mises à disposition de chercheurs *via* le CASD.

Concrètement, les données sont constituées par appariement de quatre bases de données :

- Le Fichier historique des demandeurs d'emploi (FH) qui rassemble pour chaque demandeur d'emploi des informations sur ses épisodes d'inscription à Pôle emploi sur 10 années glissantes.
- La base I-Milo (issue du système d'information des missions locales), qui rassemble pour chaque jeune suivi en mission locale des informations sur les contacts avec la mission locale et les dispositifs suivis en mission locale, depuis 2016.

- La base régionalisée des stagiaires de la formation professionnelle (Brest), qui rassemble pour chaque personne en recherche d'emploi, stagiaire de la formation professionnelle1, les caractéristiques des formations suivies depuis 2017.
- La base Mouvements de main-d'œuvre (MMO), qui rassemble pour chaque salarié du privé les informations sur les contrats de travail, depuis 2017. Les données sont issues de la Déclaration sociale nominative (DSN) qui remplacent depuis 2017 la Déclaration annuelle de données sociales unifiée (DADS-U).

Toutes ces bases de données ne disposent pas du NIR : ainsi les bases I-MILO et BREST ne contiennent « que » les nom/prénom/date de naissance. De ce fait, les méthodes d'appariement sont multiples.

2.1. Mise en œuvre du NIR haché : appariement des bases MMO et FH

Les bases de données principales MMO et FH disposent du NIR, leur appariement peut suivre le protocole dédié à l'utilisation du NIR haché dans le domaine de la recherche (pour rappel, le NIR haché dédié à la statistique publique ne peut circuler qu'au sein du SSP, Pôle emploi en est exclu).
Le tiers de confiance gestionnaire des identités est l'entreprise DATASTORM, filiale du GENES, qui est en charge de réaliser le chiffrement du NIR permettant de garantir l'unicité du résultat de l'opération (Figure 4).

En dernier lieu, le CASD joue le rôle de tiers de confiance chargé de l'appariement. Il réceptionne d'une part les identifiants (dont le NIR haché) et les variables d'intérêt constitutives des deux bases de données.

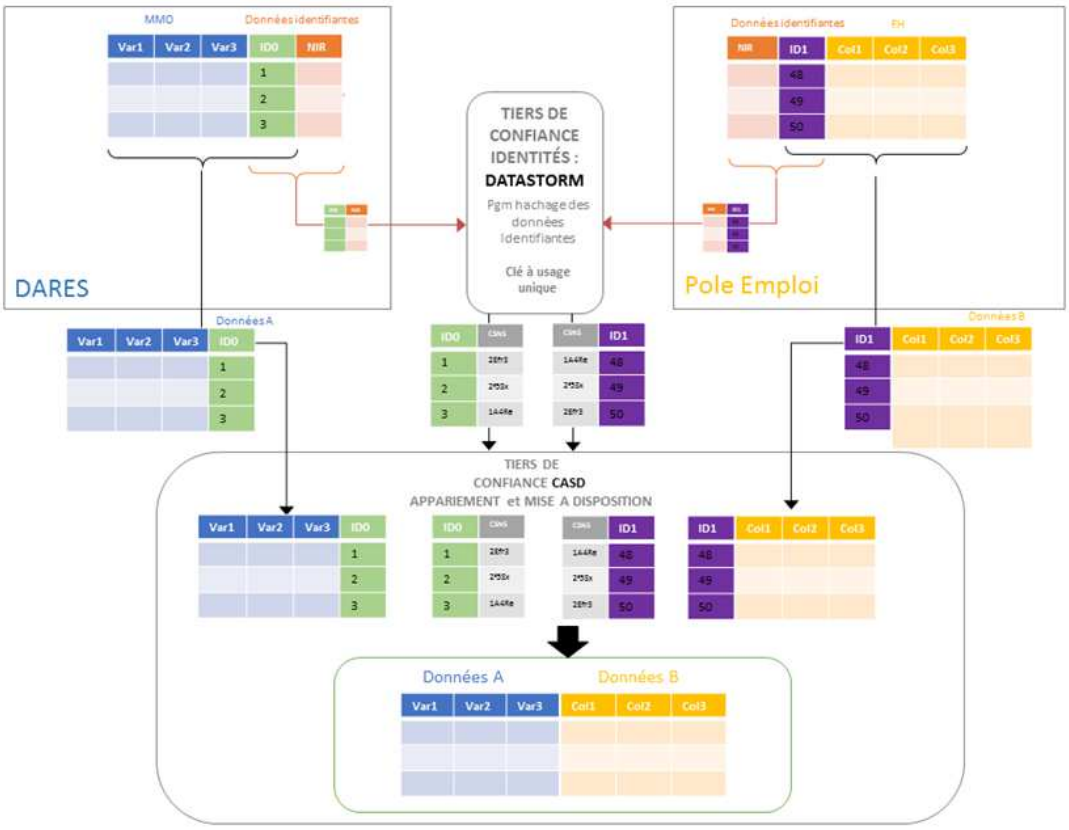

Figure 4 : appariement des bases MMO et FH via le NIR haché (CRNS)

2.2. Les appariements sur les autres variables identifiantes

Les bases de données I-MILO et BREST ne contiennent pas le NIR mais seulement le tryptique nom/prénom/date de naissance. Ces variables sont également présentes dans les bases de données

FH et MMO. Apparier sur d'autres variables que le NIR implique nécessairement un appariement imparfait (taux de matching inférieur à 100%) :

- D'une part, reconstituer le NIR correctement impliquerait de disposer, en plus des nom/prénom et date de naissance, du lieu de naissance ; cela rajouterait également une étape pour reconstituer le NIR via l'INSEE ou la CNAV, seuls à disposer de registres ;
- D'autre part, il existe des homonymes, la date de naissance n'est pas forcément complète et l'orthographe des noms/prénom pas toujours identique d'une base de donnée à l'autre.

L'appariement n'étant pas parfait, on utilise une méthode impliquant une fonction de distance entre les différents identifiants. Le choix d'un seuil de distance permet de faire varier le taux d'appariement souhaité. Plus ce seuil est élevé, plus les appariements sont robustes mais moins ceux-ci sont nombreux. Il convient dès lors de réaliser empiriquement des tests afin de choisir au mieux ce seuil. La méthode utilisée est celle de la distance de Jaro-Winkler (encadré). L'appariement est effectué en suivant les étapes suivantes :

- Les noms et prénoms sont traités pour enlever les caractères spéciaux et espaces ;
- Association des individus des deux bases avec date de naissance, nom et premier prénom identiques ;
- Pour les individus non appariés, avec date de naissance identique, comparaison des concaténations des noms et prénoms en calculant la distance de Jaro-Winkler ;
- Élimination des individus avec une distance inférieure à 0,85
- Enfin, conservation des individus présentant des noms et prénoms assez proches.

Encadré : la méthode de Jaro-Winkler

La méthode part de la distance de Jaro :

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{|m|}\right)$$

où $|s_i|$ est la longueur de la chaîne de caractères s_i ;
m est le nombre de caractères correspondants ;
t est le nombre de transpositions.

Deux caractères identiques de s_1 et de s_2 sont considérés comme correspondants si leur éloignement (i.e. la différence entre leurs positions dans leurs chaînes respectives) ne dépasse pas :

$$D = \left[\frac{max(|s_1|, |s_2|)}{2}\right] - 1$$

Le nombre de transpositions est obtenu en comparant le i-ème caractère correspondant de s_1 avec le i-ème caractère correspondant de s_2 . Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.

Winkler améliore la mesure de cette distance et utilise un coefficient de préfixe p qui favorise les chaînes commençant par un préfixe de longueur l (avec $l \leq 4$). En considérant deux chaînes de caractères s_1 et s_2 , leur distance de Jaro-Winkler d_w est :

$$d_w = d_j + (lp(1-d_j))$$

où d_j est la distance de Jaro entre s_1 et s_2
l est la longueur du préfixe commun (maximum 4 caractères)
p est un coefficient qui permet de favoriser les chaînes avec un préfixe commun.

Exemple de mesure :
Prenons deux personnels du CASD : YACINE et YASSINE

On trouve les valeurs :

$m=4$ (Y.A.I.N.E)

$|s_1|=6$

$|s_2|=5$

$t=0$ (pas de caractères dits correspondants et transposés)

$D=(max\ (6,5))/2-1=2$

La distance de Jaro est : $d_j=1/3(4/6+4/5+1)=0,82$

Et celle de Jaro-Winkler (avec un préfixe de longueur $l=2$ et un coefficient de préfixe $p=0,1$) :

$d_w=0,82+(2\times0,1\times(1-0,82))=0,86$

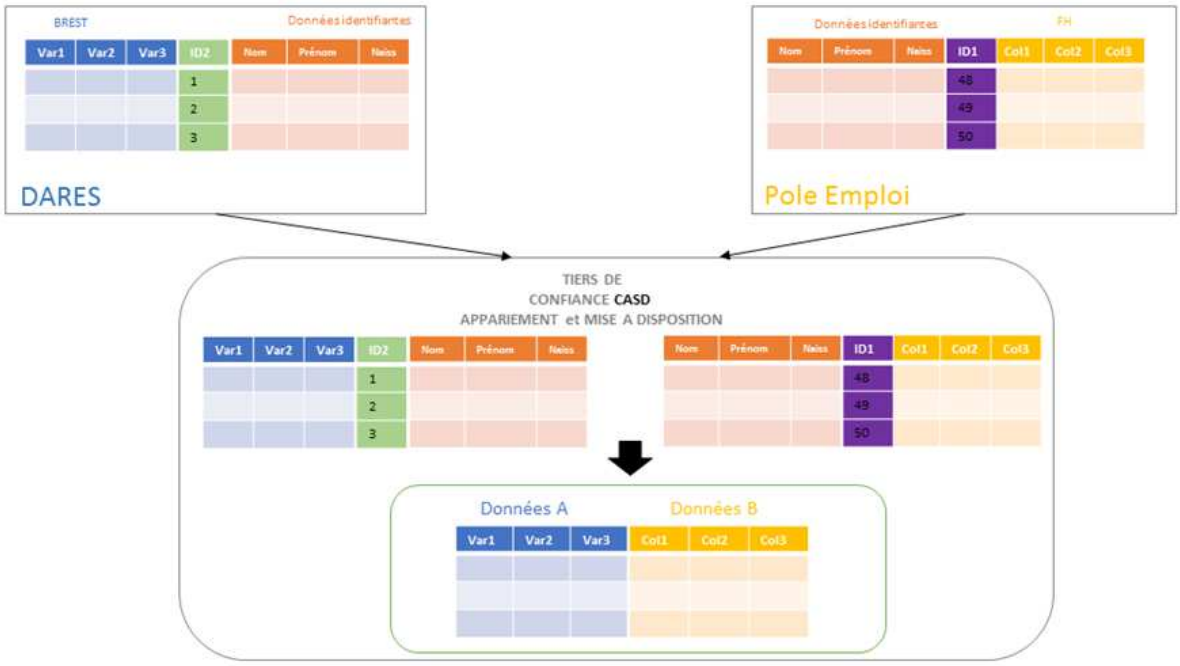

Figure 5 : appariements sur les nom/prénom/date de naissance

Au final, les appariements sur les nom/prénoms/date de naissance sont les suivants :

- FH-I-MILO
- FH-BREST
- MMO-I-MILO
- MMO-BREST
- BREST-I-MILO

En sus de ces appariements, les identifiants d'état civil sont conservés afin de permettre d'autres jonctions avec des bases de données locales (celles des conseils régionaux ou des CAF notamment) pour les projets de recherche sélectionnés dans le cadre de l'évaluation du plan d'investissement dans les compétences. Actuellement, une vingtaine de projets de recherche ont accès aux bases FORCE par l'intermédiaire du CASD.

Conclusion

On le voit, les appariements sur le NIR, s'ils ont été rendus possibles pour la recherche scientifique par les lois récentes, n'en restent pas moins des dispositifs particulièrement encadrés. Toutefois, l'expérience acquise montre que le circuit de données lié aux appariements peut être respecté sans difficulté et assez rapidement, dès lors que les tiers de confiance sont bien identifiés et que toutes les parties se sont mises d'accord sur l'objectif poursuivi.

D’autres problèmes se posent, liés à la qualité de l’information recueillie, notamment lorsque le NIR n’est pas présent dans les bases de données ou qu’il est mal codé, nécessitant de passer par l’état civil, plus ou moins complet, et qui ne garantit pas une très bonne qualité de l’appariement. Il faut toutefois garder en tête que ces appariements se font – pour l’instant – sur des populations très larges et que, avec un taux d’appariement qui reste élevé, la précision des estimations reste importante, pour peu que l’on associe la nouvelle base constituée à de nouveaux jeux de pondération.

L’obtention du CSNS pour la statistique publique pose le même type de difficulté dès lors que le NIR n’est pas disponible, mais la méthode consiste alors à ré-identifier les personnes via le RNIPP pour en tirer ensuite un CSNS. Là encore les résultats ne sont pas garantis à 100% mais restent très élevés.

Bibliographie

[1] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données)
[2] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés
[3] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
[4] Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche
[5] Jaro, M. A., « Advances in record linking methodology as applied to the 1985 census of Tampa Florida », *Journal of the American Statistical Society*, vol. 84, no 406, 1989, pp. 414-420
[6] Winkler, W. E., « The state of record linkage and current research problems », *Statistics of Income Division, Internal Revenue Service Publication* R99/04, 1999
[7] Winkler, W. E., « Overview of Record Linkage and Current Research Directions », *Research Report Series*, RRS, 2006