Le Code statistique non signifiant (CSNS) : un service pour faciliter les appariements de fichiers

Yves-Laurent Bénichou[*], *Lionel Espinasse*[**] *et Séverine Gilles*[***]

Les appariements de fichiers permettent d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux. Le Code statistique non signifiant (CSNS) a été défini par la loi pour une République numérique de 2016 afin de permettre la mise en œuvre d'appariements de fichiers à des fins statistiques en limitant l'usage du NIR (ou numéro de sécurité sociale), et garantir ainsi un niveau élevé de protection des données à caractère personnel. Le principe général est d'utiliser une clé d'appariement calculée à partir d'un chiffrement irréversible du numéro de sécurité sociale. Ce nouveau service offert par l'Insee aux organismes du service statistique public s'applique à une grande diversité de fichiers administratifs ou issus d'enquêtes. Une méthode innovante a été conçue pour identifier des personnes à partir de leurs traits d'identité. À l'issue du processus, la fiabilité de l'identification est mesurée par des indicateurs de qualité.

Les premières utilisations sont prometteuses. Le CSNS permet, par exemple, de contribuer à l'analyse de l'insertion des jeunes diplômés en facilitant l'appariement des données du système éducatif et du ministère du Travail. Il aide aussi à mesurer l'impact de la transition écologique selon les catégories de ménage en rapprochant les données du répertoire de véhicules et les informations sur les revenus.

File matching considerably increases the possibilities of studying economic and social phenomena. The Non-SignifIcant Statistical Code (CSNS) was defined by the Law for a Digital Republic of 2016 in order to allow the implementation of file matching for statistical purposes without using the NIR (or national insurance number), thus ensuring a high level of personal data protection. The general principle is to use a matching key calculated from an irreversible encryption of the national insurance number. This new service offered by INSEE to official statistical services applies to a wide variety of administrative or survey files. An innovative method has been developed to identify people on the basis of their identity. At the end of the process, the reliability of the identification is measured by quality indicators.

The first uses are promising. For example, the CSNS can contribute to the analysis of the integration of young graduates by facilitating the matching of data from the education system and the Ministry of Labour. It also helps to measure the impact of the ecological transition according to household categories by matching vehicle register data and income information.

* Expert en Data science, Unité SSP Lab, Insee
yves-laurent.benichou@insee.fr
** Adjoint à la cheffe du département de la Démographie, DSDS, Insee,
lionel.espinasse@insee.fr
*** À la date de la rédaction de l'article, cheffe de projet statistique (CSNS), DSDS, Insee
severine.gilles@insee.fr

Avec le développement de systèmes d'information performants dans de nombreux secteurs, les appariements[1] de fichiers deviennent un mode d'enrichissement des données très puissant. Ils permettent de mettre en relation des informations variées qui se trouvent dans des univers différents, et sont moins coûteux que de réaliser des enquêtes sur le terrain. Le Code statistique non signifiant (CSNS) a été défini par la loi pour une République numérique de 2016[2] pour permettre la mise en œuvre de ces appariements tout en préservant la confidentialité, en limitant l'usage du NIR[3].

> ***En reliant des informations collectées par des organismes différents, les appariements de fichiers permettent ainsi d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux.***

Deux exemples peuvent illustrer les enrichissements permis par les appariements : pour appréhender le devenir des étudiants du supérieur après leur formation, il est pertinent de croiser les données relatives à l'enseignement, détenues par le ministère de l'Enseignement supérieur avec les données d'emploi détenues par le ministère du Travail. Ou dans un autre registre, mettre en regard les données sur le parc de véhicules automobiles et leur consommation d'énergie avec le revenu de leurs propriétaires permet de déterminer l'impact de la hausse des prix des carburants sur les différentes catégories de population. En reliant des informations collectées par des organismes différents, les appariements de fichiers permettent ainsi d'accroître considérablement les possibilités d'études des phénomènes économiques et sociaux.

Pour apparier des fichiers, la situation la plus favorable est celle où l'on dispose d'un identifiant pour chaque individu, commun aux deux fichiers[4]. On peut ainsi associer les enregistrements correspondant aux mêmes individus et regrouper alors l'information les concernant issue des deux fichiers. Dans l'idéal, cet identifiant a un statut formel, est conservé dans un référentiel et permet de caractériser chaque individu sans ambiguïté et de manière unique. Le NIR, numéro d'identification au répertoire national d'identification des personnes physiques (RNIPP)[5] communément appelé « numéro de sécurité sociale » est un très bon exemple d'identifiant *(Espinasse et Roux, 2022)*.

Mais encore faut-il respecter certaines règles et notamment les attentes des citoyens relatives à la protection de leurs données personnelles. Avant 2018, pour répondre à cette attente, les appariements de fichiers sur la base du NIR pour les besoins de la statistique publique nécessitaient en particulier, pour chaque traitement, un décret en Conseil d'État après avis publié et motivé de la Commission nationale de l'informatique et des libertés (Cnil)[6]. Face à la croissance de la demande de données appariées pour les besoins

1 Les appariements constituent des interconnexions de sources de données statistiques tierces, fondées sur les données à caractère personnel et permettent de créer de nouveaux fichiers comprenant tout ou partie des variables de chacun des fichiers des sources d'origine.
2 Voir les fondements juridiques en fin d'article.
3 Le NIR est le numéro d'inscription au Répertoire national d'identification des personnes physiques (RNIPP), il est plus communément appelé « numéro de sécurité sociale ».
4 Il existe d'autres méthodes, notamment en rapprochant les traits d'identité *(voir infra),* mais celles-ci sont plus complexes à mettre en œuvre dans un processus standardisé et industrialisé.
5 Voir les fondements juridiques en fin d'article.
6 Voir les fondements juridiques en fin d'article.

de conception et d'évaluation des politiques publiques, il est devenu nécessaire d'imaginer une procédure plus simple juridiquement qui respecte le principe de minimisation des données et le statut particulier conféré au NIR par la loi Informatique et Libertés. L'idée a ainsi été émise de ne plus utiliser directement le NIR comme identifiant d'appariement, mais un identifiant non signifiant, le « NIR haché » conservant les propriétés techniques d'un identifiant tout en rendant impossible de revenir directement à l'identité des personnes.

▶ Une innovation de la loi pour une République numérique : le CSNS, un service rendu par l'Insee

> ***Le CSNS est un service de l'Insee ayant pour finalité de faciliter les appariements de fichiers de personnes physiques au sein du service statistique public (SSP), tout en garantissant un niveau de protection des données à caractère personnel plus élevé qu'avec l'usage du NIR.***

Dans ce contexte, un dispositif spécifique a été mis en place avec la création du Code statistique non signifiant (CSNS)[7]. Au-delà d'un simple code, le CSNS est aussi un service de l'Insee ayant pour finalité de faciliter les appariements de fichiers de personnes physiques au sein du service statistique public (SSP) garantissant un niveau de protection des données à caractère personnel plus élevé qu'avec l'usage du NIR.

La loi pour une République numérique précise que les institutions appartenant au service statistique public peuvent bénéficier du CSNS. Ainsi, ce service rendu par l'Insee s'adresse aussi aux Services statistiques ministériels (SSM). Les appariements peuvent être réalisés entre les fichiers de deux SSM, d'un SSM et de l'Insee ou de deux unités de l'Insee. Ces configurations peuvent être facilement étendues à trois ou quatre partenaires.

La prestation de l'Insee comporte une dimension technique pour l'identification des personnes et le chiffrement irréversible du NIR et une dimension organisationnelle avec la mise à disposition d'une application permettant aux utilisateurs de déposer leurs demandes et récupérer leurs résultats.

Tous les traitements réalisés sont enregistrés dans les registres des activités de traitement prévus par l'article 30 du Règlement général sur la protection des données (RGPD)[8] et doivent être rendus publics. Chaque traitement doit également être communiqué au Conseil national de l'information statistique (Cnis) qui pourra évaluer ce nouveau dispositif. Cette liste des traitements est diffusée publiquement sur le site du Cnis.

Le principe de minimisation des données fait aussi l'objet d'une attention particulière : seules les informations strictement nécessaires au calcul du CSNS sont échangées entre les organismes partenaires et l'Insee, puis l'Insee détruit sans délai les données qui lui ont été confiées dès que ce calcul est achevé.

7 Le CSNS est prescrit par l'article 34 de la loi pour une République numérique du 7 octobre 2016 et ses modalités opérationnelles sont précisées par un décret en Conseil d'État (n°2016-1930 du 28 décembre 2016) et par un arrêté du ministre chargé de l'économie (du 28 septembre 2020) *(voir Fondements juridiques).*

8 Voir les fondements juridiques en fin d'article.

Chiffrer le numéro de sécurité sociale : le CSNS en tant que code

Le CSNS a vocation à servir de multiples usages dans des domaines variés. Il ne s'agit pas de limiter son utilisation à l'appariement de quelques fichiers spécifiques ; il doit pouvoir être utilisé par tous les services statistiques ministériels, pour toutes leurs sources de données. Il prend ainsi appui sur un référentiel de population qui couvre toute la population vivant en France, le Répertoire national d'identification des personnes physiques (RNIPP) qui centralise l'ensemble des numéros de sécurité sociale.

Le principe général du CSNS est d'appliquer au numéro de sécurité sociale (NIR), une opération de chiffrement irréversible pour obtenir une clé d'appariement en garantissant l'impossibilité d'identifier individuellement les personnes concernées.

Chaque personne ayant un NIR unique, le calcul du CSNS donnera toujours le même résultat quel que soit le fichier sur lequel il est appliqué et permettra ainsi des appariements sans que l'on ait besoin de connaître l'identité des personnes.

Enfin, le calcul du CSNS repose sur un processus entièrement automatisé. Ce principe a été posé dès la conception du projet. Les propriétaires de fichiers ont accès à une application dédiée et sont autonomes pour faire leurs demandes et récupérer leurs résultats. Ils obtiennent des fichiers de CSNS produits sans intervention humaine, même pour le traitement des cas les plus complexes (voir *infra*). Cette automatisation présente l'avantage de rendre le coût de fonctionnement supportable pour l'Insee et de pouvoir ainsi rendre ce service gratuit, mais aussi de gagner en temps de traitement. En contrepartie, elle implique une standardisation du processus d'identification des personnes et du processus d'évaluation de la qualité des résultats. Elle conduit aussi à ce que le travail de préparation des données, propre aux spécificités de chaque fichier, relève de la responsabilité des demandeurs.

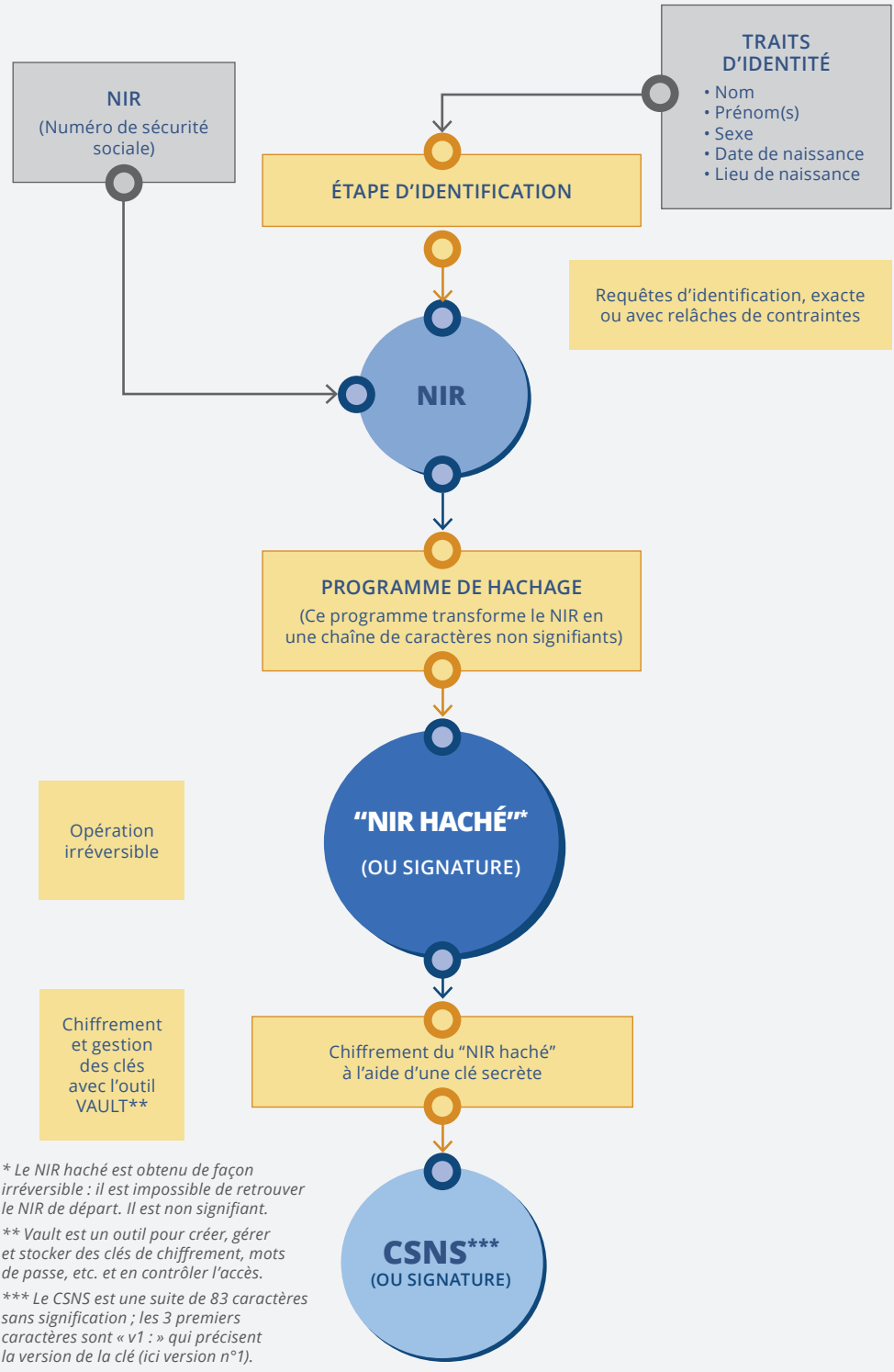
Deux possibilités pour s'adapter aux besoins

La richesse et la qualité des informations sur les identités des personnes sont variables selon les fichiers à apparier. Certains disposent déjà du NIR, d'autres ne l'ont pas mais rassemblent des traits d'identité comme le nom, les prénoms, le sexe, la date et le lieu de naissance. Deux **possibilités** sont alors offertes ***(figure 1)***. Le calcul du CSNS peut se faire soit à partir du NIR, soit à partir des traits d'identité.

Dans le premier cas, le calcul se résume simplement à chiffrer le NIR de manière irréversible. Dans le deuxième cas, le plus fréquent, une étape supplémentaire est nécessaire. Il faut d'abord retrouver le NIR des personnes à partir de leurs traits d'identité. Cette phase est appelée « étape d'identification ». Il suffit ensuite de chiffrer les NIR ainsi retrouvés. Cette deuxième façon de procéder est plus complexe à mettre en œuvre et restitue des résultats avec des niveaux de fiabilité qui dépendent de la qualité des données d'identité en entrée.

Le fonctionnement est le suivant : chaque propriétaire de fichier transmet au service CSNS de l'Insee les NIR ou les traits d'identité des individus de son fichier.

Figure 1 - Comment est obtenu le CSNS ?



** Le NIR haché est obtenu de façon irréversible : il est impossible de retrouver le NIR de départ. Il est non signifiant.*

*** Vault est un outil pour créer, gérer et stocker des clés de chiffrement, mots de passe, etc. et en contrôler l'accès.*

**** Le CSNS est une suite de 83 caractères sans signification ; les 3 premiers caractères sont « v1 : » qui précisent la version de la clé (ici version n°1).*

Cette transmission s'effectue de façon sécurisée, avec une application dédiée et accessible aux seuls demandeurs habilités du service statistique public. En retour, chaque demandeur reçoit un CSNS calculé pour chacun des individus de son fichier ***(encadré 1)***. Pour un même individu, le CSNS sera toujours le même, quelle que soit la source où figure cet individu et quelle que soit l'année où le CSNS est calculé. Avec cette clé unique et pérenne[9], les partenaires peuvent ensuite apparier leur fichier.

La technique de chiffrement répond à un haut niveau d'exigence de sécurité et comporte deux étapes. Le calcul du CSNS correspond d'abord à un hachage du NIR, programme qui transforme le NIR en une chaîne de caractères non signifiants de façon irréversible, puis à un chiffrement de ce « NIR haché » à l'aide d'une clé secrète. Cette double opération permet de garantir qu'il est impossible de retrouver le NIR à partir du CSNS. Même si la clé de chiffrement était malheureusement dévoilée, elle ne permettrait de revenir qu'au NIR haché mais pas au NIR.

▶ Encadré 1 : Le CSNS en pratique

Le service de calcul du CSNS à partir du NIR est ouvert depuis octobre 2021 et celui proposant une identification sur traits d'identité depuis octobre 2022. Il est réservé au service statistique public.

Pour y avoir accès, chaque service statistique doit au préalable signer avec l'Insee un contrat de sous-traitance qui précise les droits et obligations de chacun et indique les principes de fonctionnement du processus. Ce contrat est valable 5 ans. Une fois signé, chaque service peut effectuer autant de demandes qu'il le souhaite. Le service est gratuit.

La phase de calcul du CSNS n'est toutefois qu'une des phases du processus d'appariement. Elle s'inscrit dans une démarche plus générale où deux propriétaires de fichiers se rapprochent pour apparier leurs données, exploiter le résultat de cet appariement et le diffuser. L'Insee en tant qu'opérateur du CSNS offre un service de sous-traitance à ces propriétaires de fichiers pour la production d'une clé d'appariement commune, mais n'intervient pas sur les autres aspects.

En particulier, il convient qu'au moins un des propriétaires soit responsable de traitement au sens du RGPD et réalise toutes les démarches requises à ce titre, notamment la production d'une analyse d'impact relative à la protection des données (AIPD) si nécessaire.

La démarche de mise en œuvre d'un traitement CSNS comprend alors en général six étapes :

- établissement d'une convention entre les deux propriétaires de fichiers fixant les conditions de réalisation de l'appariement de leurs données et de leur utilisation ;
- déclaration du traitement par le ou les responsables de traitement ;
- inscription du traitement dans le programme de travail transmis au Cnis ;
- demande par chaque propriétaire d'un calcul de CSNS à l'Insee et restitution par l'Insee à chaque propriétaire du résultat du calcul avec des indicateurs de qualité ;
- appariements des données par les propriétaires de fichiers selon les modalités qu'ils auront définies dans leur convention ;
- conservation de ses données par chaque propriétaire selon des règles définies par la réglementation et rappelées dans le contrat de sous-traitance.

Les modalités pratiques pour effectuer une demande de calcul de CSNS sont simples. Après vérification des habilitations, le demandeur dépose son fichier de NIR ou de traits d'identité dans une application dédiée ouverte en ligne. Puis il reçoit en retour les CSNS et leurs indicateurs de qualité. Chaque demandeur peut réitérer ses demandes autant de fois qu'il le souhaite, notamment s'il apporte des améliorations à son fichier en entrée.

9 Les CSNS calculés sont valables 10 ans à compter de 2022, sauf si une faille de sécurité est détectée. Au bout de 10 ans (ou plus tôt en cas de faille), les clés de chiffrement des NIR seront modifiées. Le processus de renouvellement produira une table de correspondance entre les nouveaux et les anciens CSNS.

Après appariement, les CSNS doivent être conservés par chaque propriétaire de manière sécurisée et isolée dans un fichier qui ne comprend aucune variable socio-démographique, aucun NIR et aucun trait d'identité. Seul un numéro d'ordre permettra à l'avenir de faire à nouveau correspondre le CSNS avec les individus, pour de nouveaux appariements. Par ailleurs, la durée de conservation doit être proportionnée aux besoins actuels et éventuellement futurs.

Lorsque le demandeur ne dispose pas du NIR mais uniquement de traits d'identité, une étape d'identification est nécessaire avant de réaliser ce processus de hachage-chiffrement du NIR.

▶ Retrouver les NIR à partir des traits d'identité : l'étape d'identification

L'opération d'identification du NIR à partir de traits d'identité est effectuée avec un moteur spécifiquement développé pour le calcul du CSNS. Il existe déjà un processus d'identification au RNIPP, mais celui-ci est utilisé pour des besoins à finalité administrative (pour les services fiscaux par exemple). Dans ces configurations qui induisent des conséquences administratives sur la vie des personnes, aucune erreur n'est admise. L'identification n'est confirmée que si la correspondance entre les traits d'identité à vérifier et les traits d'identité du RNIPP est certaine. De ce fait, l'identification échoue parfois ; la recherche de l'exactitude de la correspondance se solde par un taux d'échec d'appariement parfois relativement élevé.

> ***Les besoins statistiques sont différents des besoins administratifs.***

Or, les besoins statistiques sont différents des besoins administratifs. Sur une population de plusieurs milliers ou centaines de milliers de personnes, quelques erreurs d'identification n'auront pas de conséquences importantes sur les résultats statistiques finaux (et aucune conséquence sur les individus eux-mêmes). Il peut donc être intéressant d'accepter quelques approximations dans la correspondance des traits d'identité si cela permet d'augmenter le taux d'identification, sous contrôle d'un taux d'erreur acceptable mais faible. Tout l'enjeu est ainsi de trouver le point d'équilibre entre maximiser l'identification et minimiser les erreurs. Ceci est d'autant plus important lorsque les sources statistiques utilisées sont des enquêtes. Les personnes enquêtées renseignent en général moins scrupuleusement leurs données d'identité dans une enquête que dans un formulaire administratif. A fortiori, les processus de lecture optique de questionnaires papier, comme dans le recensement de la population, peuvent ajouter de l'incertitude sur la qualité de la saisie des données d'identité. C'est pour cela qu'un moteur d'identification spécifique a été développé pour le CSNS.

La construction de ce moteur d'identification s'est appuyée en partie sur la théorie des appariements *(cf. Christen, 2012 ; Fellegi et al. 2014)*, mais en tenant compte de la spécificité du référentiel de population utilisé. La phase de préparation des données relève de la responsabilité de l'organisme demandeur et non de l'équipe CSNS de l'Insee, même si quelques actions sont nécessaires en début de processus de calcul afin d'adapter au mieux les données au fonctionnement du moteur.

Ensuite, le très gros volume du référentiel de population (le RNIPP comprend 130 millions d'occurrences) a conduit à concevoir trois étapes pour l'identification d'un NIR avec le souci d'optimiser les temps de traitement selon la difficulté des cas à traiter. Pour cela, le principe de traiter les cas simples avec des processus peu gourmands en temps, et de réserver les processus complexes aux cas qui le nécessitent vraiment, a été retenu.

Un enchaînement d'étapes, de la plus simple à la plus complexe

Un enchaînement de trois étapes est conçu pour retrouver un NIR, adaptées aux différents niveaux de complexité de la recherche. ***(figure 2)***.

La première étape est une requête dite « exacte ». Les éléments d'identification (nom, prénoms, date de naissance, code géographique du lieu de naissance) sont recherchés dans le RNIPP, de façon exacte. L'ensemble des prénoms doit être exact également. À cette étape, l'identification suit un principe fondamental : elle ne peut avoir lieu que si un seul écho[10] est candidat à l'identification. À noter également que les recherches sont aussi effectuées sur les anciens noms (lorsque les personnes en ont changé) ou les noms d'usage (marital souvent lorsqu'il figure au RNIPP).

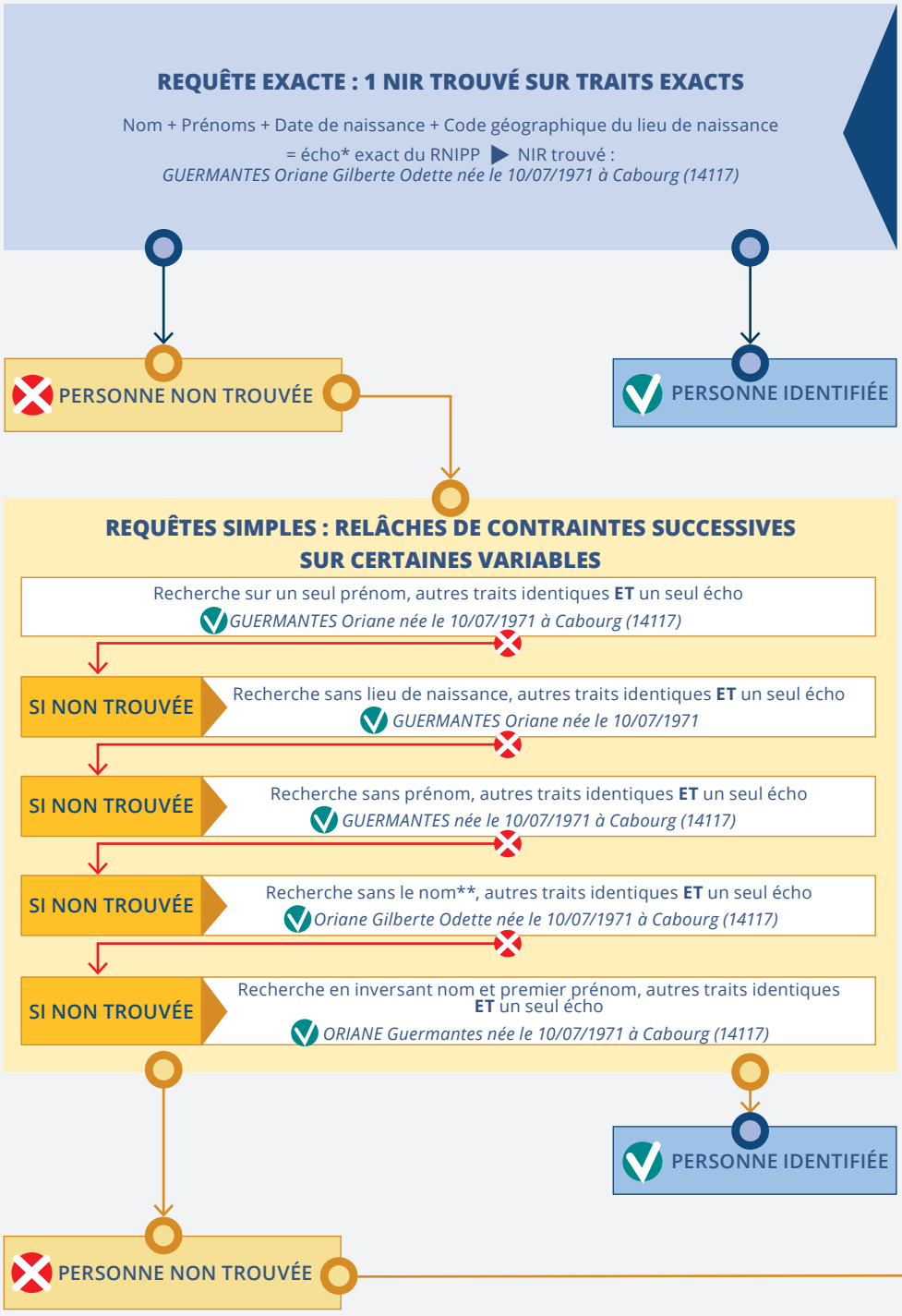
Vient ensuite une deuxième étape dite de « requêtes simples ». On autorise alors quelques « relâches de contraintes »[11] sur certaines variables. Là encore, ne sont retenus que les échos sans concurrent. Cinq assouplissements successifs des contraintes d'identification sont mis en œuvre :

- une relâche simple sur les prénoms : les éléments d'identification sont tous identiques sauf les prénoms pour lesquels on autorise une identification sur un seul et non sur tous ;
- une relâche simple sur le code géographique du lieu de naissance : les éléments d'identification sont tous identiques, à l'exception du code géographique du lieu de naissance ; pour les prénoms, l'exactitude est exigée sur le premier ;
- une relâche simple et totale sur les prénoms : les éléments d'identification sont tous identiques, à l'exception des prénoms ;
- une relâche simple et totale sur le nom : les éléments d'identification sont tous identiques à l'exception du nom ; pour les prénoms, l'exactitude est exigée sur le premier prénom. Cette relâche sur le nom peut paraître surprenante au premier abord, le nom semblant être l'élément déterminant d'une identité. Mais en fait, les tests ont montré que de nombreux échecs d'identification étaient dus à l'usage du nom marital dans les fichiers des utilisateurs alors que celui-ci pouvait être inconnu du RNIPP (le nom marital n'est pas une variable obligatoire du RNIPP). Cette requête permet ainsi d'identifier de nombreuses femmes mariées qui ne déclarent pas leur nom de naissance dans les enquêtes mais leur nom d'usage ;

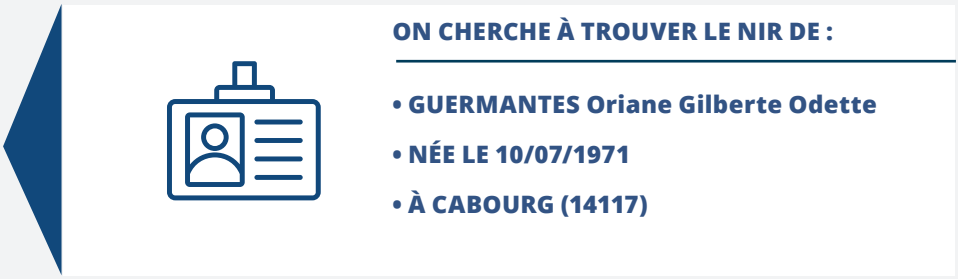
10 Un écho est un retour d'information à la suite d'une recherche sur les traits d'identité d'une personne. Si les traits d'identité recherchés sont incomplets ou très fréquents, plusieurs personnes peuvent correspondre et dans ce cas, plusieurs échos peuvent être restitués.

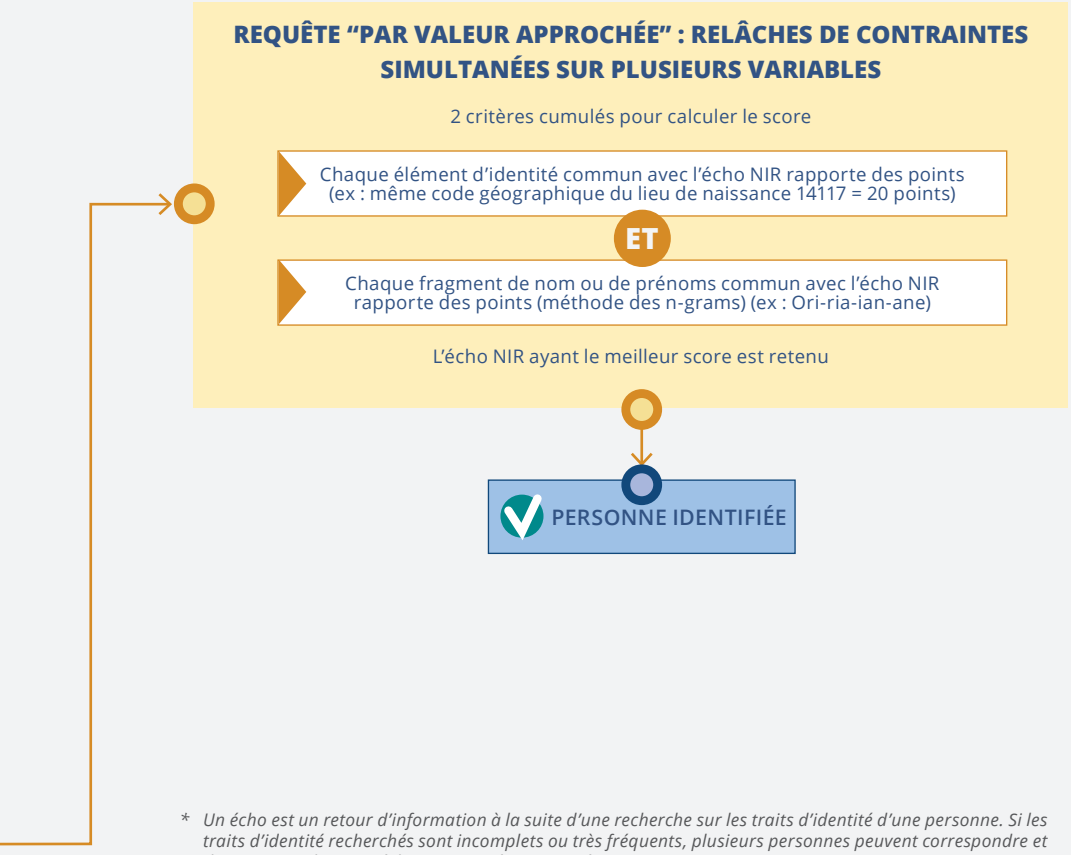
11 « Relâcher des contraintes » : réduire le nombre de composantes concordantes de l'identité nécessaires à la validation d'une identification.

Figure 2 - Exemple (fictif) pour retrouver un NIR (ou numéro de sécurité sociale)



à partir des traits d'identité





* *Un écho est un retour d'information à la suite d'une recherche sur les traits d'identité d'une personne. Si les traits d'identité recherchés sont incomplets ou très fréquents, plusieurs personnes peuvent correspondre et dans ce cas, plusieurs échos peuvent être restitués.*

** *Cette requête permet de retrouver les personnes qui se déclarent avec leur nom marital.*

- une identification par interversion des nom et premier prénom : les éléments d'identification sont tous identiques, à l'exception près que les champs nom et premier prénom sont intervertis pour la recherche d'identification.

Ces relâches successives sur les variables permettent en général de compléter significativement l'identification de l'étape de requête exacte. Pour les fichiers administratifs testés et avec une bonne qualité de remplissage des traits d'identité, l'identification sur l'ensemble des deux premières étapes permet en général de retrouver plus de 95 % des individus. En revanche, pour les fichiers d'enquêtes où la qualité de remplissage des traits d'identité est moins bonne, ces deux étapes donnent des résultats moins complets (80 %).

> ***Pour les fichiers administratifs testés et avec une bonne qualité de remplissage des traits d'identité, l'identification sur l'ensemble des deux premières étapes permet en général de retrouver plus de 95 % des individus.***

Après ces deux premières étapes, il reste toutefois à identifier les cas les plus complexes. Certains ont été délibérément exclus de la sélection des requêtes précédentes, leurs données d'identités étant sujettes à caution : il s'agit des personnes déclarées nées les 1er janvier et 31 décembre (ces dates étant parfois des dates « par défaut » indiquées lorsque l'information exacte n'est pas connue[12]) ou avec une date de naissance (jour/mois) à (00/00). Sont également exclues des sélections précédentes les identifications aboutissant à une personne de sexe différent de celui déclaré. Les autres cas complexes correspondent simplement à toutes les personnes non identifiées précédemment.

▶ Une méthode innovante pour les cas les plus complexes

Pour traiter ces cas, une troisième et dernière requête, « par valeur approchée », permet de relâcher des contraintes sur plusieurs variables de façon simultanée. La personne à identifier correspond à plusieurs échos dans le RNIPP et il s'agit de choisir celui qui présente le maximum de caractéristiques communes avec elle. Le choix s'appuie sur le classement des échos, réalisé en calculant un score. Le score est une somme de « points » attribués en fonction des éléments communs entre l'écho du RNIPP et la personne à identifier.

Ces scores sont d'abord attribués sur des éléments d'identité exactement identiques. Par exemple, le fait d'avoir le même code géographique du lieu de naissance donne 20 points ; le fait d'avoir les mêmes jours et mois de naissance donne 20 points ; le fait d'avoir le même nom donne 10 points ; le fait d'avoir le même premier prénom donne 10 points ; etc.

Mais cette méthode ne suffit pas pour bien discriminer les échos et éviter d'aboutir à des scores *ex æquo*. On ajoute alors un autre critère pour préciser les recherches, et calculer des scores plus fins.

12 Les personnes véritablement nées le 1er janvier ou le 31 décembre seront alors identifiées lors de la dernière étape.

Ce second critère de recherche consiste à filtrer les échos du RNIPP en fonction des fragments de noms ou de prénoms qu'ils ont en commun avec la personne à identifier. Ces fragments sont des tuiles de caractères, plus ou moins longues (3 à 5 caractères). Par exemple, le prénom Justine présente 5 tuiles de 3 caractères : *jus – ust – sti – tin – ine*. Cette technique appelée « méthode des n-grams » permet de discriminer de manière efficace les différents échos ayant des points communs avec les traits d'identité d'une personne à retrouver. Elle permet également de s'affranchir de quelques erreurs de saisie ou d'orthographe. Ensuite, chaque tuile commune apporte des points qui s'ajoutent au score. Ces opérations sont réalisées à l'aide du logiciel Elasticsearch. ***(encadré 2)***.

L'écho du RNIPP ayant obtenu le score le plus élevé est retenu et le CSNS sera alors calculé par chiffrement du NIR correspondant à cet écho comme aux étapes précédentes.

▶ Une obligation : mesurer la qualité de l'identification

Les dispositions réglementaires prévoient également que le service CSNS produise une mesure de la qualité de l'identification des personnes. Celle-ci est indispensable afin que les utilisateurs puissent apprécier la fiabilité de l'appariement à venir et adapter en conséquence d'éventuels traitements statistiques. La qualité de l'identification s'apprécie au regard de deux axes : la qualité générale du fichier des traits d'identité en entrée du processus et la qualité de l'identification par enregistrement individuel.

La qualité du fichier en entrée relève de la responsabilité du demandeur, mais le service CSNS lui fournit des informations pour l'aider à l'améliorer. Pour chacune des variables « nom, prénoms, année de naissance, jour et mois de naissance et code géographique du lieu de naissance », des indicateurs de taux d'anomalie et de taux de valeur manquante sont communiqués. Ces informations sont fournies directement par l'application mise à disposition des utilisateurs. Ceux-ci sont alors autonomes pour tester plusieurs versions de leur fichier. Après une première analyse, ils peuvent ainsi repérer les variables pour lesquelles des actions d'amélioration sont nécessaires. Il est possible de recommencer cette analyse autant de fois que nécessaire.

La qualité de l'identification de chaque enregistrement relève de la responsabilité du service CSNS. Cette qualité est appréhendée sous l'angle de la mesure des faux-positifs. Un faux-positif est une identification qui aboutit à trouver un NIR différent de celui de la personne recherchée. On se trompe donc de personne et l'appariement rapprochera les données de deux personnes différentes. À l'inverse, les faux-négatifs sont des personnes qui n'ont pas été identifiées alors qu'elles auraient dû l'être. La mesure de la qualité de l'identification individuelle de chaque enregistrement est présentée sur la base du taux de faux-positifs. Plus ce taux est faible, meilleure est la qualité supposée.

▶ Encadré 2 : Un moteur de recherche au cœur du moteur d'identification au RNIPP

Ces dernières années, la palette d'outils à disposition des statisticiens s'est considérablement enrichie : langages de programmation tels que R, Python et Julia, bibliothèques de modules très performants pour traiter le nettoyage, la transformation et l'analyse des données, *machine learning*, *deep learning,* visualisation, traitement du langage…

Ces évolutions se sont accompagnées d'avancées technologiques toutes aussi innovantes dans la façon de stocker, organiser et traiter les données : démocratisation de l'utilisation des systèmes de gestion de bases de données SQL, bases NoSQL, nouveaux formats de stockage, répartition des traitements. Autant d'avancées technologiques, majoritairement mises à disposition sous licences libres qui permettent de traiter efficacement des données massives.

Parmi ces outils, les moteurs de recherche ont largement bénéficié de toutes ces évolutions récentes, ont participé aux succès des grands acteurs de l'Internet et sont désormais incontournables dans la vie quotidienne. Nous les utilisons tous sciemment plusieurs fois par jour en ouvrant notre navigateur et tout autant inconsciemment, puisque les applications de médias sociaux, de commerce électronique, de cartographie, de transport, de diffusion de musique ou de vidéos utilisent toutes leurs capacités de recherche et d'analyse en temps réel à très grande échelle.

Pour les besoins d'identification au RNIPP où il s'agit d'apparier, avec des temps de traitements raisonnables, des fichiers de traits d'identité qui peuvent contenir jusqu'à plusieurs millions de lignes avec le répertoire des personnes qui lui en contient plus de 130 millions, l'équipe de projet CSNS s'est rapidement orientée vers une solution utilisant le moteur de recherche Elasticsearch pour plusieurs raisons :

1) Une architecture technique performante et évolutive

– l'architecture d'*Elasticsearch* est distribuée : plusieurs instances (ou nœuds) peuvent être lancées sur un ensemble de serveurs (ou *ferme de serveurs*) et collaborer. Les données, selon leur volume, peuvent être découpées en plusieurs partitions puis distribuées et répliquées sur les différents nœuds afin d'assurer performance et sécurité grâce aux dispositifs de répartition de charge et de haute disponibilité.

En cas de défaillance d'un ou plusieurs nœuds de la ferme, le système continue à fonctionner en mode dégradé, sans autre conséquence que des temps de réponse plus longs, et ceci tant que les nœuds restants peuvent accéder à au moins une version opérationnelle des données.

– l'architecture d'*Elasticsearch* est extensible : la puissance de calcul offerte par la ferme peut être adaptée à l'évolution des besoins ; il est possible d'étendre la ferme dynamiquement en ajoutant des serveurs.

2) Des fonctions de traitement et d'analyse de texte avancées

Elasticsearch offre nativement des fonctionnalités telles que la recherche en texte intégral, des analyseurs pour traiter et normaliser le texte, la recherche sur synonymes, l'analyse de données géospatiales, ou encore la segmentation en unités lexicales qui permet de découper les phrases en mots ou en n-grams de mots ou de caractères.

3) Une architecture flexible

Elasticsearch s'intègre sans difficulté dans un projet informatique, car il est facile de l'interroger à partir de tous les langages de programmation tels que Java, R ou Python pour ne citer que ceux qui figurent au schéma directeur informatique de l'Insee.

En complément de tous les mécanismes et fonctionnalités précédemment décrits, la force du moteur de recherche *ElasticSearch* est sa capacité à retrouver l'information parmi des millions de lignes en temps réel. Et pour ce faire, il "triche" un peu : tout est pré-calculé lors de la phase dite d'indexation pendant le chargement des données ; par "tout" il faut entendre les calculs de tous les n-grams possibles, les variantes des termes sans les caractères spéciaux, sans les majuscules, sans les mots vides de sens, etc. Cette phase prend beaucoup de temps (6 heures pour le RNIPP) et est réalisée une fois par mois dans le cas du CSNS pour intégrer les récentes mises à jour du RNIPP.

Puis, lorsqu'une requête de recherche est soumise, le serveur qui la reçoit la distribue sur les serveurs de la ferme et un score de pertinence pour chaque enregistrement correspondant à la requête est calculé. Ce score de pertinence est basé sur le module de similarité utilisé par Elasticsearch et paramétré par l'Insee pour évaluer les similarités entre les termes recherchés et les termes indexés. Les enregistrements avec les scores les plus élevés sont considérés comme plus pertinents et sont proposés en premier dans les résultats de recherche.

En conclusion, la solution retenue pour le CSNS permet de bénéficier de l'efficacité d'un moteur de recherche intégrant la combinaison de plusieurs facteurs : pré calcul des valeurs de tous les champs de recherche, répartition de ces informations sur plusieurs serveurs, distribution des requêtes sur tous ces serveurs pour augmenter la puissance de calcul et enfin un moteur de similarité pour produire les scores de pertinence des résultats.

> ***La stratégie consiste à minimiser les erreurs d'identification (avoir peu de faux-positifs) et indiquer de manière transparente le risque pris pour chaque enregistrement.***

La stratégie consiste à minimiser les erreurs d'identification (avoir peu de faux-positifs) et indiquer de manière transparente le risque pris pour chaque enregistrement. Le choix a ainsi été fait d'identifier tout le fichier en entrée, exceptés les quelques cas extrêmes où les données sont trop parcellaires, et d'indiquer pour chaque enregistrement une estimation de la probabilité d'avoir un faux-positif. Cette estimation prend la forme d'un indicateur de qualité en 7 modalités, allant de « parfaitement fiable » (1) à « non fiable » (7)[13].

L'utilisateur reste maître de son choix : retenir ou non l'identification et le CSNS afférent proposé. Selon ses objectifs, il peut éventuellement reconsidérer les cas qu'il juge trop incertains, soit en améliorant l'identification avec d'autres informations (par exemple l'adresse qui ne figure pas dans le processus CSNS), soit en refusant ces identifications et en traitant alors ces données avec des techniques de redressement analogues à celles du traitement de la non-réponse.

On aurait pu imaginer une autre approche consistant à livrer uniquement des CSNS pour les identifications jugées de bonne qualité. Mais la concertation préalable avec les futurs utilisateurs a mis en évidence le besoin de disposer du maximum d'informations, même de qualité moindre, pour conserver l'opportunité de les retraiter et de les améliorer, dès lors qu'une évaluation du niveau de fiabilité est fournie pour chaque enregistrement.

Une mesure de la qualité adaptée aux différentes méthodes d'identification

Les modalités pratiques du calcul des indicateurs de qualité doivent tenir compte de deux contraintes. D'une part, les méthodes d'identification sont différentes selon les étapes du processus, allant d'une identification exacte à une identification par valeur approchée. D'autre part, le calcul réel d'un taux de faux-positifs requiert que l'on dispose des NIR des personnes à identifier à comparer avec les NIR trouvés par le moteur. Or, ces fichiers ne sont pas très nombreux et ne constituent pas la majorité des fichiers pour lesquels le service CSNS sera utilisé. Par ailleurs, si un utilisateur dispose du NIR, le calcul des CSNS sera réalisé par simple hachage-chiffrement sans passer par l'étape d'identification sur traits d'identité.

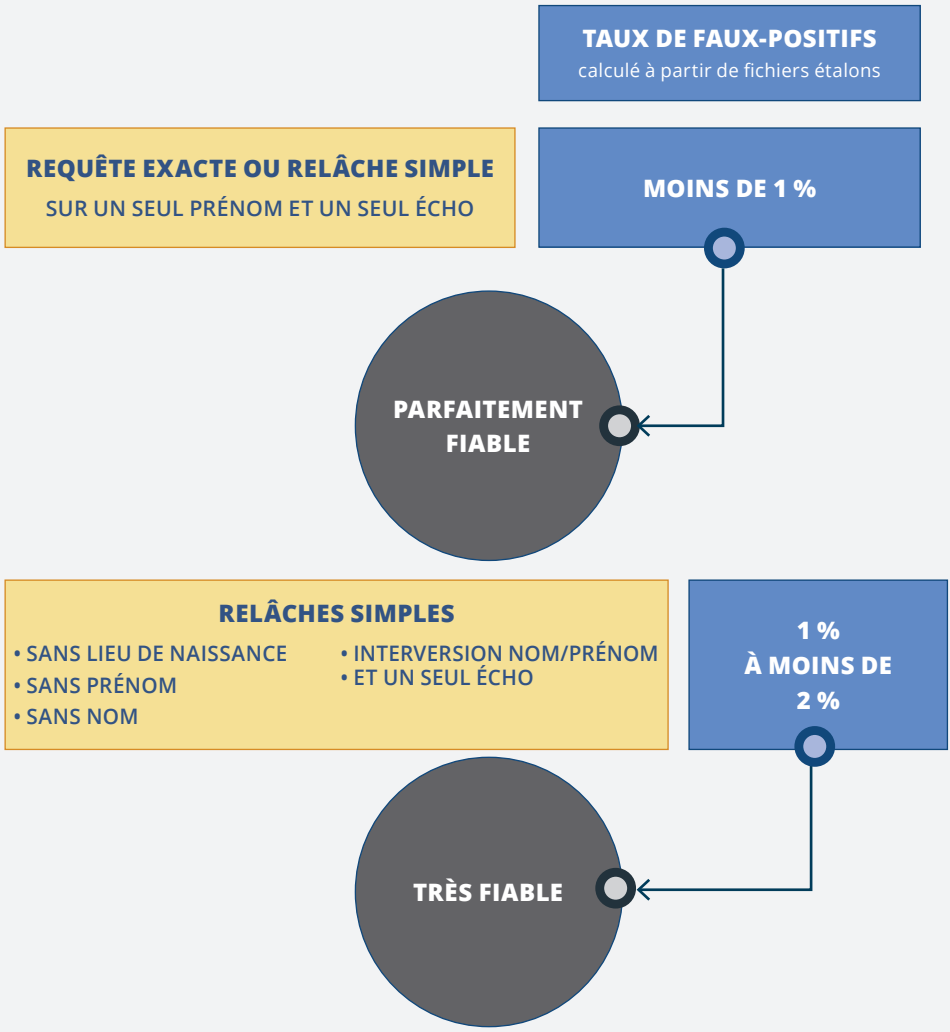
Lors de la phase de projet du CSNS, un étalonnage des taux de faux-positifs a ainsi été établi avec des fichiers comprenant le NIR pour définir des taux moyens par étape du processus. Par exemple, les calculs ont été réalisés sur les Déclarations sociales nominatives (DSN) et les NIR retrouvés après identification des traits d'identité ont été comparés aux NIR réels de ces fichiers, permettant ainsi de calculer des taux de faux-positifs sur un gros volume de données. De tels calculs empiriques ont pu aussi être menés sur l'Échantillon démographique permanent (EDP), sur les enquêtes annuelles de recensement dont le NIR avait été trouvé par un dispositif extérieur au CSNS *(Jabot et Treyens, 2018)*, et sur des fichiers

13 Parfaitement fiable (1) / Très fiable (2) / Fiable (3) / Assez fiable (4) / Peu fiable (5) / Très peu fiable (6) / Non fiable (7).

de la DREES[14] relatifs aux dispositifs d'insertion. Les valeurs moyennes de taux de faux-positifs calculées lors de cet étalonnage ont été prises comme référence pour déterminer le niveau de qualité de chaque étape.

Avec cette méthode, il ressort que les taux de faux-positifs moyens varient de 0 à 2 % lorsque la personne a été identifiée par les requêtes exactes ou simples (deux premières étapes) ***(figure 3)***.

Figure 3 - Critère principal de qualité, le taux de faux-positifs : plus il est faible, meilleure est la qualité.



14 Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), service statistique ministériel dans les domaines de la santé et du social.

En revanche, pour la méthode de la requête par valeur approchée (dernière étape), la mesure de la qualité doit faire l'objet d'une approche spécifique, même si le principe de recherche des faux-positifs reste maintenu. Le score calculé pour chaque individu constitue une information intéressante, mais celle-ci ne peut pas être utilisée directement. En effet, avec la méthode des n-grams, plus un mot est long, plus il contient de tuiles de caractères et plus son score sera potentiellement élevé. Il est alors impossible de déterminer une liaison directe entre valeur du score et probabilité d'avoir un faux-positif. Toutefois, des scores très faibles correspondent souvent à des faux-positifs. Il s'agit alors de combiner cette information avec une autre.

Le rapport entre le score de l'écho retenu (le meilleur) et celui venant immédiatement en deuxième position est alors apparu comme une autre information à exploiter. En effet, les tests ont montré que plus l'écart entre les deux premiers échos est faible, plus la probabilité de se tromper est forte. Autrement dit, si les deux meilleurs échos trouvés pour une personne ont un score proche, il est probable que la différence de proximité avec les traits d'identité originaux soit trop faible pour être significative. Cette forte proximité fait prendre le risque de se tromper de personne et de retenir un faux-positif. Ainsi, les niveaux de faux-positifs calculés empiriquement sur les mêmes fichiers que pour les étapes précédentes ont été classés selon une double échelle de valeurs : celle du niveau de score, et celle du rapport « score du 2e écho sur score du 1er écho ».

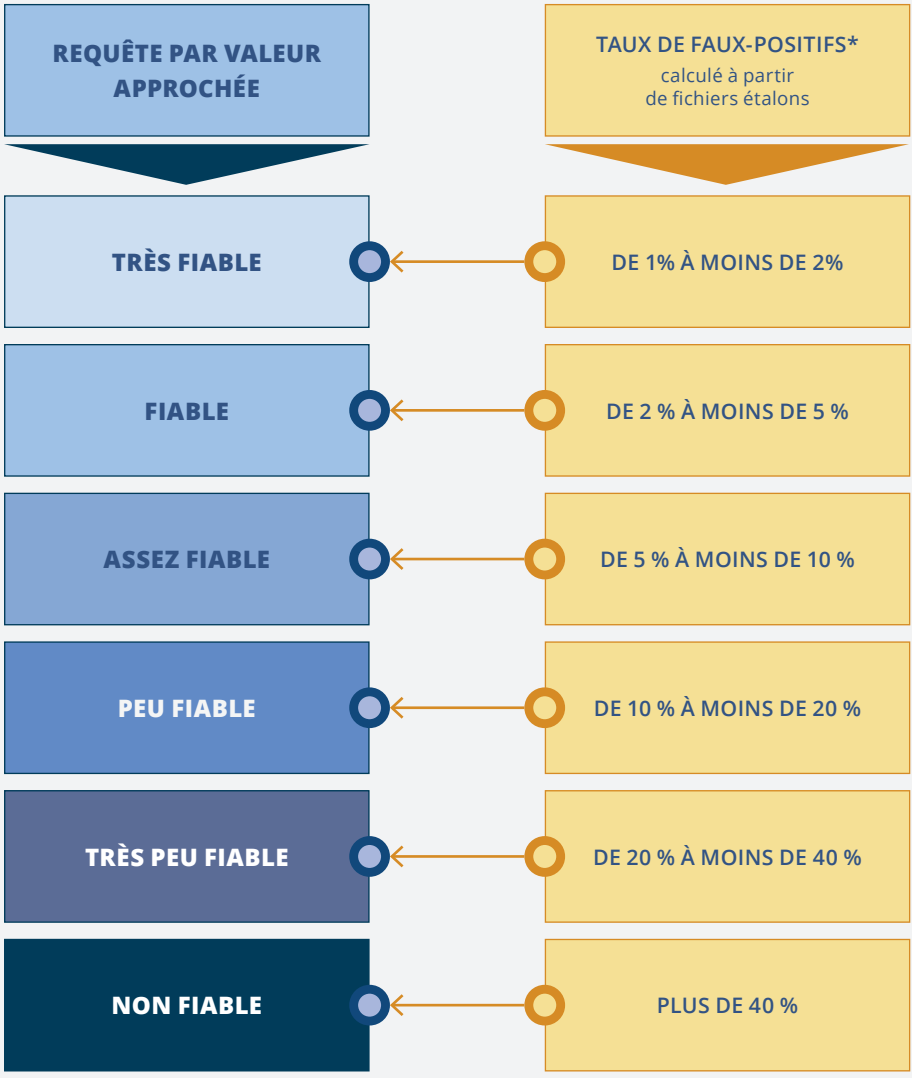
Au final, l'indicateur de qualité en 7 modalités, allant de « parfaitement fiable » à « non fiable » est calculé comme suit : les identifications trouvées à l'étape exacte et à la première requête simple de relâche sur le prénom sont classées en « parfaitement fiable » avec une probabilité de taux de faux-positifs de moins de 1 %. Celles trouvées aux autres requêtes simples sont classées en « très fiables » avec une probabilité de taux de faux-positifs de 1 à 2 %. Les identifications de l'étape par valeur approchée ont des indicateurs de « très fiable » à « non fiable » selon leur classement au regard de la valeur de leur score et de la valeur du ratio entre le deuxième et le premier score de la recherche *(Malherbe, 2022)* ***(figure 4)***.

Dernière étape de la mesure de la qualité : l'évaluation sur l'ensemble du fichier. Cette dernière phase informe l'utilisateur de la part des enregistrements pour chacun des sept niveaux de qualité : ces informations sont disponibles de façon automatique dans l'application dédiée.

En complément, d'autres outils sont fournis aux utilisateurs pour les aider à appréhender la qualité de l'identification de leur fichier. En particulier, une répartition par âge est calculée sur la population initiale du fichier en entrée du processus et sur la population identifiée en sortie. La comparaison des deux permet de voir si une sous-population particulière est sur-représentée parmi les échecs d'identification. Des informations analogues sont fournies pour comparer les populations nées en France et celles nées à l'étranger, la qualité des données d'état civil pouvant parfois être différente.

Ces indicateurs de qualité ont été testés avec des SSM volontaires et ils se sont révélés utiles et pertinents pour les appariements à venir.

▶ Figure 4 - Critère principal de qualité, le taux de faux-positifs : Davantage de modalités pour la requête par valeur approchée.



** corrélé à la valeur du score du 1er écho*
▶ plus elle est élevée, meilleure est l'identification

et

à l'écart de score entre le 1er écho et le 2e (mesuré par le ratio des scores)
▶ plus cet écart est faible, plus il est difficile de les départager et plus le risque de se tromper d'identification est élevé.

▶ Des premières utilisations prometteuses

En facilitant les appariements entre différentes sources, le CSNS contribue à l'extension des possibilités d'analyse des phénomènes économiques et sociaux. Le gisement est considérable et les premières utilisations donnent une idée du potentiel résultant de ce nouveau processus.

Lors de la phase de projet du CSNS, quatre services statistiques ministériels ont participé activement à de nombreux tests (Dares, Drees, SDES et SIES)[15], notamment pour faire des propositions, et évaluer la robustesse des choix méthodologiques et la pertinence des calculs d'indicateurs de qualité. Les exemples qui suivent montrent la diversité des sujets traités et l'intérêt en termes de connaissance de la société et d'évaluation des politiques publiques.

> ***Mieux mesurer l'insertion des diplômés de l'enseignement supérieur.***

Mieux mesurer l'insertion des diplômés de l'enseignement supérieur, c'est l'objectif du projet InserSup mené par le SIES avec la Dares. L'objectif est de produire des indicateurs d'insertion professionnelle, par établissement formateur et diplôme, sur l'ensemble des diplômes, et de mettre cette information à disposition des élèves et étudiants sur ParcourSup, MonMaster, Affelnet, Onisep[16], etc. pour les aider à choisir leur formation. Il s'agit aussi d'informer les acteurs territoriaux et les employeurs sur le lien formation-emploi, et plus généralement d'éclairer le débat public sur l'insertion professionnelle.

Le CSNS contribue à ce projet en facilitant l'appariement de sources administratives de provenances diverses sans avoir à recourir à des enquêtes spécifiques. Les sources mobilisées sont les différents répertoires étudiants du SIES, et la déclaration sociale nominative DSN (Dares) qui comprend de nombreuses informations sur les salariés. Est également envisagée l'utilisation de la base des non-salariés, des fichiers du recensement de la population et des fichiers fiscaux. Le service rendu par le CSNS permet ainsi de multiplier les appariements possibles entre les différentes sources d'observation, d'écourter les délais de mise à disposition et d'offrir une information exhaustive sur le champ couvert.

Plus généralement, l'obtention du CSNS pour les 40 millions de salariés du Système d'information statistique sur les mouvements de main d'œuvre (Sismmo) de la Dares permettra notamment d'étudier l'emploi étudiant en l'appariant avec les bases des inscrits de l'enseignement supérieur.

15 La Direction de l'animation de la recherche, des études et des statistiques (Dares) est le service statistique ministériel dans le domaine du travail ; la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) est le service statistique ministériel dans les domaines de la santé et du social ; le Service des données et études statistiques (Sdes) est le service statistique ministériel dans le domaine du logement, de la construction, des transports, de l'énergie, de l'environnement et du développement durable ; la Sous-direction des systèmes d'information et des études statistiques (Sies) est le service statistique ministériel de l'enseignement supérieur et de la recherche.

16 Parcoursup est une plateforme Web destinée à recueillir et gérer les vœux d'affectation des futurs étudiants de l'enseignement supérieur français ; MonMaster est une plateforme de consultation de l'intégralité des diplômes de Master et de dépôt des candidatures en 1re année de Master ; la procédure Affelnet permet d'affecter les élèves de troisième dans les lycées de leur académie ; l'Onisep est un opérateur de l'État qui produit et diffuse toute l'information sur les formations et les métiers.

Dans un tout autre domaine, le CSNS pourra contribuer à mieux mesurer l'impact de la transition écologique sur les différentes catégories de ménages. En rapprochant les données du Répertoire statistique des véhicules routiers (RSVERO) avec les données sur les revenus, la localisation exacte et les caractéristiques des ménages issues de la source Fidéli[17], c'est un nouveau champ d'analyse qui s'ouvre ; par exemple, le lien entre les caractéristiques des véhicules (puissance, type de carburant) et le revenu de leur propriétaire pourra être mieux appréhendé. Ces informations sont essentielles pour définir ou évaluer les politiques publiques relatives à la précarité énergétique ou à la transition écologique.

Dans le champ social, le CSNS va permettre d'enrichir les données des enquêtes sur l'autonomie et la dépendance. L'appariement de l'enquête CARE (enquête capacités, aides et ressources des seniors) avec les informations sur les prestations telles que l'APA (allocation personnalisée d'autonomie) et l'ASH (aide sociale à l'hébergement) permettra de suivre l'évolution de la dépendance des seniors deux ans après l'enquête. Parallèlement, le rapprochement des données de l'enquête VQS (vie quotidienne et santé) avec les données des régimes sociaux et celles sur l'insertion et l'emploi (notamment la déclaration sociale nominative) permettra de relier la prise en charge de la perte d'autonomie avec l'insertion professionnelle, et de traiter la question des incapacités en fin de carrière. À plus long terme, le CSNS ouvrira également des perspectives d'études et d'analyses dans de nombreux domaines tels que le devenir des enfants confiés à l'Aide sociale à l'enfance, les parcours des bénéficiaires du revenu de solidarité active (RSA) ou de l'APA… Sur ces sujets, l'exhaustivité des sources administratives appariées permettra de disposer de données territorialisées et régulièrement actualisées.

> ***Toutes ces nouvelles possibilités d'analyse offertes par le CSNS peuvent intéresser le monde de la recherche.***

Toutes ces nouvelles possibilités d'analyse offertes par le CSNS peuvent intéresser le monde de la recherche, au-delà du cercle restreint du service statistique public (SSP) *(Gadouche, 2019)*. Ainsi, même si le processus de calcul du CSNS est réservé au SSP, les résultats finaux des appariements peuvent être mis à disposition plus largement, mais sans faire figurer le CSNS lui-même.

Par ailleurs, la loi pour une République numérique de 2016 a aussi prévu un dispositif spécifique pour le monde de la recherche[18]. Une même opération de chiffrement du numéro de sécurité sociale pour aboutir à un code non signifiant utilisable comme clé d'appariement devient une nouvelle possibilité offerte aux chercheurs. La différence avec le CSNS est toutefois que ce code de recherche est attaché à un projet de recherche en particulier et ne peut pas être utilisé pour un autre projet. Ainsi un individu avec un code non signifiant dans un projet de recherche n'aura pas le même code dans un autre projet de recherche.

Quelques mois seulement après l'ouverture complète du service CSNS en octobre 2022, ce nouveau dispositif est déjà utilisé par cinq services statistiques ministériels et par plusieurs unités de l'Insee. L'intégration systématique et annuelle du CSNS est programmée pour plusieurs fichiers très utilisés par la statistique publique : Fideli

17 Fichiers démographiques sur les logements et les individus.
18 Articles 7 à 9 du décret n°2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche (voir Fondements juridiques).

(revenus fiscaux), les enquêtes annuelles de recensement (EAR), les fichiers issus de la Déclaration sociale nominative (DSN), l'échantillon démographique permanent (EDP)… L'expression « CSNSiser un fichier » se popularise auprès des statisticiens. Même si elle n'est pas très élégante, elle augure de la mise en place de bonnes habitudes qui rendront plus fluides et plus faciles les appariements à venir, et contribueront à augmenter encore et encore les sources de données nécessaires à une juste observation de notre société.

▶ Fondements juridiques

- 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, Article 30). In : *Journal officiel des Communautés européennes*. [en ligne]. Mis à jour le 04 mai 2016. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A32016R0679*.

- Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés modifiée et en particulier son article 30. In : *site de la CNIL*. [en ligne]. Mis à jour le 14 mars 2021. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.cnil.fr/fr/la-loi-informatique-et-libertes#article30*.

- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique et en particulier son article 34. In : *site de Légifrance.* [en ligne]. Mis à jour le 08 octobre 2016. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746*.

- Décret n° 82-103 du 22 janvier 1982 relatif au répertoire national d'identification des personnes physiques. In : *site de Légifrance.* [en ligne]. Mis à jour le 01 juillet 2021. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000520382*.

- Décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. In : *site de Légifrance*. [en ligne]. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033735139*.

- Arrêté du 28 septembre 2020 pris en application des articles 3 et 4 du décret n° 2016-1930 du 28 décembre 2016 portant simplification des formalités préalables relatives à des traitements à finalité statistique ou de recherche. In : *site de Légifrance*. [en ligne]. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000042382996*.

▶ Bibliographie

- CHRISTEN Peter, 2012. *Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://link.springer.com/book/10.1007/978-3-642-31164-2*.

- ESPINASSE Lionel et ROUX Valérie, 2022. Le Répertoire national d'identification des personnes physiques (RNIPP) au cœur de la vie administrative française. In : *Courrier des statistiques*. [en ligne]. Novembre 2022. Insee. N°8, pp.72-92. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/6665188?sommaire=6665196*.

- FELLEGI, Ivan, SUNTER, Alan, JARO, 2014*. Approach to Record Linkage (Method)*. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://cros-legacy.ec.europa.eu/content/fellegi-sunter-and-jaro-approach-record-linkage-method_en*.

- GADOUCHE Kamel, 2019. Le centre d'accès sécurisé aux données (CASD), un service pour la *data science* et la recherche scientifique. In : *Courrier des statistiques*. [en ligne]. Décembre 2019. Insee. N°N3,pp.76-92. [Consulté le 14/03/2023]. Disponible à l'adresse : *https://www.insee.fr/fr/information/4254227?sommaire=4254170*.

- JABOT P. et TREYENS P.E. (2018). Proposition d'un nouvel appariement de l'enquête CARE par identification du plus proche écho. Actes des journées de méthodologie statistique 2018. [Consulté le 14/03/2023]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_PPT_TREYENS_JMS2018.pdf*.

- MALHERBE Lucas, 2022. Méthodologie des appariements individuels. JMS 2022. [Consulté le 14/03/2023]. Disponible à l'adresse : *http://jms-insee.fr/jms2022s07_1/*.