Appariement de données pseudonymisées

Maxence Guesdon [1⋆°], Catherine Quantin [2⋆⊗], Eric Benzenine [3⋆],

⋆ *CHRU Dijon, Service de Biostatistique et d'Informatique Médicale (DIM), Dijon, F-21000, France ; Université de Bourgogne, Dijon, F-21000, France*

⊗ *INSERM, CIC 1432, Dijon, France ; Dijon University Hospital, Clinical Investigation Center, clinical epidemiology/ clinical trials unit, Dijon, France*

° *Institut National de Recherche en Informatique et Automatique, France*

Résumé

Les fichiers administratifs de la France, notamment médicaux et sociaux, offrent un potentiel d'études statistiques très important. Le NIR est largement utilisé dans les administrations des secteurs médico-sociaux, permettant théoriquement de nombreux croisements, sous réserve du respect de la réglementation.
Cependant, la loi interdit la manipulation de données médicales non anynonymes, rendant difficile la réalisation d'études nécessitant plusieurs sources de données sociales et médicales.
Mais depuis les années 70, de nouvelles techniques informatiques sont apparues et ont été utilisées depuis les années 90 sur des données personnelles. Ainsi, le hachage permet de pseudonymiser des données sans possibilité de retour à la donnée originale, permettant un certain niveau d'anonymat. D'autre part, les techniques de chiffrement permettent d'assurer notamment la confidentialité des données, par des systèmes basés sur des clés.
Ces nouvelles techniques peuvent être combinées pour permettre la réalisation d'appariements sécurisés de fichiers anonymisés. L'utilisation de ces techniques peut ainsi être généralisée pour faciliter les recherches et la réalisation d'études statistiques publiques. Il est souhaitable que la gestion de la confidentialité soit confiée à une Autorité de gestion des clés.

Abstract

Administrative records in France, especially medical and social records, have huge potential for statistical studies. The NIR is widely used in medico-social administrations, and this would theoretically provide considerable scope for data matching, on condition that the legislation on such matters was respected.

1. maxence.guesdon@inria.fr
2. catherine.quantin@chu-dijon.fr
3. eric.benzenine@chu-dijon.fr

The law, however, forbids the processing of non-anonymized medical data, thus making it difficult to carry out studies that require several sources of social and medical data.
Since the 70s, new computer techniques came into being and these have been applied since the 90s on personal data. Hashing can be used to deidentify data with no possibility of returning to the original information, thus ensuring a certain degree of anonymity, and encryption techniques can be used to ensure, in particular, data confidentiality using key-based systems.
These new techniques can be combined to safely link anonymized files. The use of these techniques could also be generalized to facilitate research and to carry out public statistical studies. The management of this confidentiality should, however, be entrusted to a 'Key-Management' Authority.

Mots-clés

Appariement, données personnelles, hachage, chiffrement, anonymat.

Introduction

Au début des années 1970, la France disposait d'un potentiel statistique très prometteur avec ses riches fichiers administratifs mobilisables ensemble ou associés à une enquête.
En 1974, l'informatisation du répertoire de l'état civil sous le nom très maladroit de projet SAFARI suscita une vive émotion dans l'opinion publique, avec la crainte d'un fichage généralisé de la population et de potentielles dérives en cas de gouvernement totalitaire. Ce numéro national d'identité permettait, en effet, de croiser les informations issues des multiples fichiers administratifs relatives aux mêmes personnes comme le font actuellement et avec précaution les instituts de statistique de tous les pays nordiques.
Ce débat a alimenté une vaste réflexion sur les mesures permettant de protéger la vie privée et les libertés face au développement de l'informatique. Il a donné naissance à la Loi Informatique et Libertés votée le 6 janvier 1978, instituant la CNIL [4].
Cette loi a été globalement très positive ; elle a cependant verrouillé la statistique administrative, soit en limitant la statistique dans un traitement à plat fichier par fichier, soit en imposant des procédures démesurément lourdes et historiquement fortement dissuasives (comme la prise d'un décret en Conseil d'Etat).
Dans le souci d'éviter les dangers d'un croisement général des fichiers administratifs, la CNIL a opté pour une stratégie de cantonnement, c'est-à-dire l'impossibilité d'utiliser le même identifiant dans tous les fichiers. Comme le NIR était déjà largement répandu dans les administrations sociales et de santé, la CNIL a limité l'usage du NIR aux secteurs du travail, de la santé et des institutions sociales. Pour les autre secteurs (finances [5], éducation, ...), de nouveaux identifiants dits "sectoriels" ont donc été créés, sans possibilité d'un lien entre le NIR et ces nouveaux identifiants, lien qui aurait pu être utilisé à des fins de statistique publique ou de recherche. Aujourd'hui, les techniques d'appariements sécurisés permettent de dépasser ces limites. C'est l'objet de cette communication.

4. Commission Nationale de l'Informatique et des Libertés

5. Cette situation a évolué. Ainsi, l'administration fiscale associe désormais le NIR à son identifiant sectoriel, pour supprimer les doublons et transmettre aux organismes sociaux les informations fiscales utiles à leur gestion, par exemple lorsqu'une allocation est soumise à un plafond de revenus.

Par ailleurs, la loi interdit la manipulation de données personnelles médicales non anonymes. Cependant, l'appariement de fichiers de données médicales personnelles est impossible si ces données sont anonymes au sens strict, puisqu'alors plus aucune information ne permet de relier un individu dans deux fichiers différents. Dans la suite, nous utiliserons le terme d'anonymisation pour parler d'une anonymisation relative dans ce sens, basée notamment sur la pseudonymisation. La pseudonymisation consiste à remplacer systématiquement chaque valeur d'un champ identifiant par une autre valeur, sans pouvoir revenir à la valeur initiale.
Notre communication, limitée à la pseudonymisation de l'identifiant, ne traite pas de l'identification indirecte, d'autant plus probable et à risque qu'on démultiplie les données par rapprochement de fichiers. Ce risque d'identification dépend bien sûr de l'information, certes inconnue, détenue par le tiers mal intentionné. On a donc affaire à un "niveau d'anonymisation", ou une anonymisation partielle plus ou moins importante, niveau qui n'est pas vraiment prévu par la loi.
Depuis les années 70, les techniques informatiques ont également évolué. Les techniques actuelles permettent ainsi, en préservant l'anonymat, de réaliser des appariements inter-institutionnels sécurisés grâce au hachage (cf. section 1) de l'identifiant commun ou, en l'absence de ce dernier, par appariement probabiliste (cf. section 2.2).
Dans la section 4, nous rappelons la stratégie proposée pour l'épidémiologie par Catherine Quantin dans [18] pour dépasser de façon sécurisée les contraintes précédentes et nous proposons d'étendre cette stratégie à la statistique publique pour lui offrir une flexibilité élargie sans perdre en sécurité.

1 Hachage

1.1 Principes

Les techniques de hachage[23] sont des procédures informatiques consistant à calculer une *empreinte* (ou signature) de taille fixe à partir de données de n'importe quelle taille.
Une fonction de hachage permet donc de plonger n'importe quelle donnée en entrée dans un ensemble fini, dont le cardinal est très grand [6]. Cela signifie qu'il n'y a pas d'opération inverse qui, à partir de l'empreinte, permettrait de retrouver la donnée initiale, puisqu'il existe une infinité de données ayant la même empreinte.
De plus, la *distance* entre deux empreintes de deux données est indépendante de la distance entre ces deux données. Une différence minime entre deux données en entrée aboutit à deux empreintes très différentes (principe dit de "l'effet avalanche"). Au contraire, deux données très différentes en entrée peuvent avoir des empreintes proches ou identiques.
Par exemple, le hachage des chaînes "Dupont" et "Dupond" par la fonction SHA256 donne les empreintes suivantes :

```
SHA256("Dupont") = 3bde3a5999601d8fa7b6bcc6bfdd2ee6a9fb473043d9768fbf8274b5936ef4d2
SHA256("Dupond") = 535a7594e59be910df06483d24371c7697854fa84d8ed8c0f400126edc25af3a
```

Une bonne fonction de hachage présente un *risque de collision faible*, c'est-à-dire que pour des données différentes mais de taille de même ordre, la probabilité d'avoir la même empreinte est extrèmement faible [7]. Les collisions qui seraient introduites par le hachage sont

6. Par exemple, le nombre de signatures différentes produites par SHA-256 est de 2^{256}, soit un nombre supérieur à 10^{77}.

7. Par exemple, pour des mots de 80 bits (10 caractères codés sur un octet chacun), le risque de collision par SHA256 est de l'ordre de 10^{-31}.

minimes par rapport aux problèmes d'homonymies que l'on peut rencontrer en pratique lorsqu'on manipule les données en clair sur les noms, prénoms, dates de naissances, ...
Puisqu'une fonction de hachage n'a pas de fonction inverse, on dit que le traitement par hachage est *irréversible*. Cependant, si on connaît la fonction de hachage, il est possible de retrouver une donnée en entrée à partir de sa signature, grâce à des attaques dites *par dictionnaire*.

1.2 Résistance aux attaques

Le principe de ces attaques est le suivant. Si on connaît la fonction de hachage utilisée, on peut l'appliquer à un ensemble de chaînes de caractères. On peut alors construire une table de correspondance entre chaque chaîne de caractères et son empreinte issue du hachage.
Les taux de collisions (*i.e.* deux chaînes différentes donnent la même empreinte) des algorithmes de hachage sont extrêmement faibles. Aussi, ayant connaissance d'une empreinte en particulier, il suffit de se référer à la table de correspondance pour identifier de manière quasi certaine la chaîne de caractères initiale.
Ce type d'attaque pose donc un problème de confidentialité des données si on utilise le hachage pour anonymiser (ou plutôt pseudonymiser) des données personnelles.
La solution consiste donc à modifier la chaîne avant de lui appliquer la fonction de hachage. Une façon classique de procéder consiste en l'ajout d'un *sel*, c'est-à-dire une clé secrète qu'on ajoute à chaque donnée avant de calculer son empreinte. Si notre clé est par exemple `XZ!#45`, on ajoutera cette clé en début ou fin de la chaîne à hacher :

```
SHA256("DupontXZ!#45") = cd0c6a7852dc50474778d2599a6bf85d5c8c1f31a6c4e348a52e4fcd04b8d660
SHA256("DupondXZ!#45") = 7e20b3c86d4c1508f1c4b7650ffa62e3fd379bb10fad9b3c618449cb9088d0d0
```

Sans connaître la taille de la clé ni son contenu, les attaques par dictionnaire deviennent en pratique impossibles, puisque, même en faisant l'hypothèse que la taille de la clé est comprise entre 1 et 20 octets [8], il faudrait alors construire $256 + 256^2 + 256^3 + ... + 256^{20}$ tables de correspondance, ce qui pose un problème de temps de calcul et d'espace de stockage.
Une autre façon de procéder est d'appliquer une fonction à la chaîne, fonction secrète ou utilisant une clé secrète permettant soit de modifier la chaîne à hacher, soit de calculer un sel différent pour chaque chaîne [9].
Il existe plusieurs fonctions de hachage "standard"[12] (MD5, SHA1, SHA256, SHA512, ...), faisant continuellement l'objet de recherches pour éprouver leur résistance aux attaques. Ainsi, il est maintenant recommandé par l'ANSSI[1] d'utiliser la méthode SHA-256.

1.3 Utilisation pour l'appariement

Les techniques de hachage permettent, en les appliquant sur les données identifiantes, de pseudonymiser les fichiers à apparier. Cependant, cette pseudonymisation présente des inconvénients.

8. 1 octet = 8 bits = 256 possibilités
9. C'est la méthode utilisée dans le logiciel ANONYMAT, développé au CHU de Dijon et ayant été validé par la CNIL pour l'anonymisation de données à des fins d'appariements[16].

En effet, la moindre erreur de saisie dans un nom aboutira à une signature complètement différente pour ce nom par rapport à la signature obtenue sur le nom correct. Des traitements de normalisation en amont peuvent limiter ces problèmes [17](SOUNDEX, passage en minuscules, suppression des accents, ...). Toujours pour la même raison, il n'est pas possible d'utiliser un calcul de distance d'édition (par exemple les distances de Levenshtein, de Hamming, ...[2]) pour s'en servir dans des appariements déterministes [10] (cf. section 2.1).
Le hachage est déjà utilisé pour l'appariement de données provenant de plusieurs fichiers afin de conserver un anonymat relatif[21].
Signalons également l'utilisation d'un double hachage lorsque les fichiers à apparier proviennent par exemple de plusieurs établissements. Il est nécessaire de hacher de la même façon (avec la même clé secrète) les champs identifiants dans tous les fichiers, afin de permettre l'appariement selon ces champs. Cependant, l'établissement recueillant ces fichiers procède à un second hachage (avec une seconde clé secrète) afin de rendre anonymes les données agrégées vis-à-vis des établissements ayant produit les fichiers. [16].
Enfin, du fait du caractère irréversible du hachage, il convient de conserver les données non hachées, sous peine de ne pouvoir les exploiter qu'avec d'autres données ayant subi le même hachage. Cela implique une gestion des clés utilisées, avec par exemple une clé par étude, voire l'existence d'une Autorité de gestion des clés (cf. section 4.3).

2 Appariements déterministe et probabiliste

Il existe deux types d'appariements, selon les données à apparier. On pourra lire utilement [3] pour une revue des méthodes d'appariement et leur utilisation sur des données de santé.

2.1 Appariement déterministe

L'appariement dit déterministe consiste à déterminer les champs identifiants dans les deux sources de données à apparier, puis à définir une mesure de distance et un seuil à partir duquel deux enregistrements sont considérés comme correspondant au même individu. Le terme "déterministe" vient du fait que les seuils choisis ne dépendent pas des données à apparier, c'est-à-dire que les même seuils sont utilisés même si des données supplémentaires sont ajoutées aux fichiers à apparier.
L'application classique de cette méthode consiste à décider d'apparier les enregistrements pour lesquels les identifiants sont strictement identiques. Ainsi, en appariant sur le NIR, ou un NIR doublement haché, une telle méthode d'appariement "strict" est facile à réaliser [11] et sa fiabilité est la même que la fiabilité du champ identifiant utilisé.
Dans cette famille de méthodes d'appariement, plusieurs raffinements sont possibles. On peut ainsi concaténer tous les champs identifiants, appliquer une mesure de distance sur cette concaténation et comparer le résultat à un seuil. On peut également appliquer une mesure de distance différente pour chaque champ pour ensuite obtenir une distance globale par pondération.

10. Il est tout de même possible, avant hachage, de découper une information par exemple en tronçons de n caractères hachés séparément, puis d'appliquer un calcul de distance sur ces tronçons hachés, la distance pouvant être fonction du nombre de tronçons hachés identiques, ou une mesure plus complexe faisant intervenir par exemple des filtres de Bloom comme dans [22].
11. Pour des fichiers mis dans une base de données de type SQL, une simple requête par jointure suffit alors pour réaliser l'appariement.

Enfin, la mesure de distance peut également être définie sur chaque champ et avec un résultat binaire (0 ou 1). On peut alors définir des règles associant une décision d'appariement à chaque configuration de similitudes et différences entre deux enregistrements.

2.2 Appariement probabiliste

L'appariement probabiliste est utile lorsqu'on ne dispose pas d'un champ identifiant sans ambiguité les individus (comme le NIR) qui soit commun aux deux sources de données à apparier, et que l'établissement de règles (par exemples à partir de distances entre champs) n'est pas possible, comme c'est notamment le cas lorsque les informations ont été anonymisées par hachage (cf. section 1.3).
La nature probabiliste de ces méthodes vient de ce qu'elles utilisent des poids associés à chaque champ utilisé comme identifiant, appelés poids unitaires. Ces poids unitaires dépendent des différentes valeurs présentes dans les champs utilisés comme identifiants, de leur fréquence, etc. Ces poids unitaires sont ensuite additionnés pour obtenir des poids composés.
Deux seuils sur ces poids composés permettent de classifier les paires d'enregistrements en "Chaînage", "Non chaînage" ou "Indécision". Ces seuils sont choisis de façon ad hoc, selon les études et les contraintes associées : précision nécessaire, nature et qualité des données, tolérance d'erreurs par défaut ou par excès, possibilités de vérification et validation, ...
Contrairement aux méthodes d'appariement déterministes, ajouter des données dans les fichiers à apparier modifiera les poids utilisés dans la décision d'appariement, et donc le choix des seuils.
Le cadre théorique de ces méthodes d'appariement probabiliste a été posé dans [5] en 1968. En 1995, Jaro applique pour la première fois ces méthodes sur des données de santé dans [10] à l'aide d'un programme informatique. En 1998, [15] décrit la première application de la méthode de Jaro à des fichiers ayant été préalablement anonymisés par hachage.

2.3 Principes

On cherche à apparier deux fichiers constitués d'enregistrements, chacun composé de plusieurs champs.
L'objectif est de rapprocher les données d'un même patient en limitant les erreurs :

- Doublons : Ne pas associer les informations du même individu (changement de nom, erreur de saisie, ...),
- Collisions : Associer à tort les informations de 2 personnes différentes.

La figure 1 illustre ces différents cas.

	même individu	individus différents
même nom	Vrai positif	Faux positif = Collision
noms différents	Faux négatif = Doublon	Vrai négatif

FIGURE 1 – Doublons et collisions

L'idée de la méthode est de prendre en compte l'information apportée par chaque valeur de chacun des champs choisis comme identifiants (nom, prénom, date de naissance, ...), et sa fréquence. Ainsi, le sexe sera beaucoup moins discriminant que la date de naissance, car il ne pourra prendre la plupart du temps que deux valeurs possibles. De même,

dans un fichier comportant uniquement des nouveaux-nés, l'année de naissance porte peu d'information.
Pour cela, un poids unitaire sera attribué à chaque caractéristique identifiante, avec une valeur positive dans le cas où deux enregistrements correspondent et une valeur négative dans le cas où deux enregistrements ne correspondent pas.
Le modèle de Fellegi et Sunter propose de répartir les paires en deux ensembles M (pour "matched", les paires qui correspondent au même individu) et U (pour "unmatched").
Pour chaque champ identifiant i, on calcule deux probabilités m_i et u_i. m_i est la probabilité que les deux enregistements aient la même valeur dans le champ i quand la paire appartient à M, u_i est la probabilité que les deux enregistements aient la même valeur dans le champ i quand la paire appartient à U.
Une fois ces deux probabilités connues, le poids unitaire associé à un champ sera $log\frac{m_i}{u_i}$ (valeur positive) quand les valeurs du champ i correspondent, sinon le poids sera $log\frac{1-m_i}{1-u_i}$ (valeur négative).
Comme on ne connaît pas les paires qui correspondent au même individu, puisque c'est le but de la méthode d'appariement, ces poids unitaires sont estimés grâce à l'algorithme EM (*Expectation-Maximization*) introduit par Winkler [25] ou une de ses variantes ultérieures [4, 3]. Ces algorithmes procèdent par itération, en se servant des données à apparier pour faire converger les estimateurs de m_i et u_i.
Une fois obtenus les poids unitaires, ces derniers peuvent être additionnés (ce sont des *log* de rapports de probabilités) pour obtenir un poids composé. La méthode donne les probabilités d'appartenance à M et à U des paires correspondant à chaque poids composé. La figure 2 donne une intuition de ces probabilités en fonction du poids composé.
On obtient donc une zone "Pas chaînage", pour laquelle la probabilité d'appartenir à M est faible, tandis que celle d'appartenir à U est grande. Dans une autre zone "Chaînage", la probabilité d'appartenir à M est grande tandis que celle d'appartenir à U est faible. Enfin, une troisième zone "Indécision" ne permettra pas de décider automatiquement de chaîner ou non les deux enregistrements en question. On aura donc deux seuils.
Selon la finalité de l'appariement, on utilisera des seuils plus ou moins élevés pour classer chaque paire d'enregistrements dans ces trois catégories.

2.4 Exemple

Illustrons cette méthode sur un appariement réalisé sur les données d'hospitalisation d'un établissement de santé (appariement de deux années successives), basé sur trois champs identifiants : le nom, le prénom et la date de naissance. Dans la suite, nous nommons A et B les deux fichiers à apparier.
Pour chaque champ, le poids unitaire n'est autre que la logvraisemblance issue de la figure 1 ; elle est donc additive pour l'ensemble des caractéristiques identifiantes de l'individu. Le calcul des poids unitaires, dépendant des données, donne les résultats de la figure 3.
Pour chaque paire d'enregistrements composée d'un enregistrement du fichier A et d'un enregistrement du fichier B, le poids composé est calculé en additionnant le poids unitaire associé à chaque champ selon l'égalité ou la différence pour ce champ entre les deux enregistrements. La figure 4 montre les poids composés de quelques configurations des égalités et différences parmi les 8 possibles ($2 \times 2 \times 2$) pour nos 3 champs.
La figure 5 montre le résultat d'une comparaison entre deux enregistrements.
Pour chaque paire, la décision de chaînage est prise en fonction des seuils de poids composé choisis selon la précision nécessaire à l'étude (cf. figure 6).

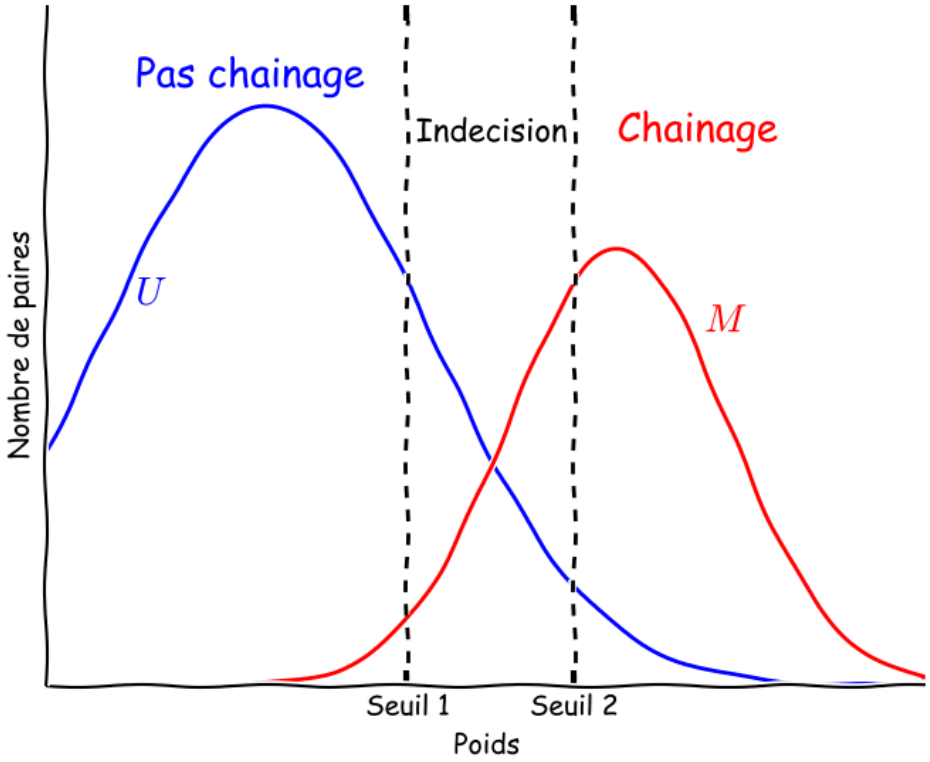

FIGURE 2 – Décision de chaînage selon le poids composé.

	Nom	Prénom	Date de naissance
Poids si égaux (1)	8.4	5.7	10.3
Poids si différents (0)	-2.8	-3.5	-3.1

FIGURE 3 – Résultats du calcul des poids unitaires pour chacun des 3 champs utilisés pour l'appariement.

Pour les paires pour lesquelles il n'y a pas de décision automatique de chaînage ou non, c'est-à-dire pour les configurations dans lesquelles le poids composé est entre les seuils 1 et 2 (cf. figure 2), une validation [12] manuelle est possible, en remontant éventuellement au dossier des patients [13]. Cette validation, permettant d'apparier ou non des enregistrements en indécision, peut se faire en partie automatiquement par des procédures supplémentaires appliquées à une partie des paires d'enregistrement, celles en dessous du seuil de chaînage automatique pour lesquels une bonne proportion (par exemple, toujours figure 6, une grande partie des 725 paires concordant selon le nom et la date de naissance devraient

12. Plutôt qu'une validation, on peut également utiliser l'information descriptive fournie. Ainsi, dans notre exemple, le modèle nous dit que sur les 725 cas en 011, 78.66% « devraient » être chaînés.

13. La validation manuelle dépend également de l'étude à faire et de l'importance qu'elle revêt. Ainsi, pour une étude épidémiologique concernant l'impact d'un médicament, on souhaitera faire toutes les vérifications nécessaires et les validations manuelles permettant d'apparier correctement le plus grand nombre d'enregistrements. Pour une étude moins critique, croisant par exemple la réussite au baccalauréat et les notes en cours d'année, on pourra se contenter d'avoir un taux d'appariement plus faible, sans aller manuellement faire les vérifications, coûteuses, nécessaires.

	Nom	Prénom	DdN	Poids composé
Sans discordance (111)	+8.4	+5.7	+10.3	+24.4
Discordance sur le nom (011)	-2.8	+5.7	+10.3	+13.2
Discordance sur la DdN (110)	+8.4	+5.7	-3.1	+11
Discordance sur tous les champs (000)	-2.8	-3.5	-3.1	-9.4

FIGURE 4 – Calcul du poids composé selon la configuration des égalités et différences.

	Nom	Prénom	Date de naissance	
	Dupont	François	29/01/1940	
	Dupont	François	29/03/1940	
Poids	+8.4	+5.7	-3.1	= 11

FIGURE 5 – Exemple de calcul de poids composé pour deux enregistrements.

être chaînées). Cette validation automatisée peut également utiliser d'autres champs discriminants supplémentaires mais non utilisés pour la classification automatique.
Cette méthode de chaînage probabiliste a notamment été utilisée pour la détermination du statut vital en croisant les données hospitalières et les données de mortalité nationales dans [6].
Conformément à la législation, les données avaient préalablement été anonymisées en utilisant la technique du hachage. En pratique, les comparaisons des champs sont donc faites sur des données hachées. Ainsi, le tableau de la figure 5 ressemble plutôt à celui de la figure 7.

2.5 Appariement mixte

La méthode d'appariement mixte utilisée par Marine Guillerm [9] est un mélange des deux méthodes précédentes. Il s'agissait pour elle d'apparier les données scolaires ou universitaires identifiées par un INE [14] correct avec le fichier des notes au bac identifié par un INE occasionnellement défaillant.
L'INE est utilisé pour réaliser un appariement seulement partiel car sa fiabilité n'est pas totale. On l'utilise donc pour compléter par un appariement probabiliste à l'aide d'une méthode par apprentissage prenant en compte cet identifiant.

2.6 Appariement par blocs

L'application des méthodes probabilistes sur des fichiers de taille même moyenne conduit tout de même, par le produit cartésien qui est fait sur les enregistrements, à des temps de calculs qui peuvent être importants.
Pour pallier ce problème, on utilise la méthode dite d'appariements par blocs (ou "blocking" [10, 7]), qui permet par exemple de ne croiser que certains ensembles d'enregistrements d'un fichier A et d'un fichier B. Ainsi, si le champ sur le sexe est fiable, on peut se contenter de ne croiser que les enregistrements qui correspondent pour ce champ. On peut faire de même en utilisant par exemple l'année de naissance. On peut également faire un blocage sur plusieurs champs à la fois (par exemple sexe et année de naissance). Enfin, on

14. Identifiant National Etudiant.

Concordance Nom	Prénom	DdN	Fréquence	Seuils	Poids	$P(m)$	$G(u)$
0	0	0	1 452 966 248		-9.4	6e-08	99.99
0	1	0	4 880 218		-0.2	5e-04	99.99
1	0	0	304 887		1.8	4e-03	99.99
0	0	1	46 081		1.4	0.04	99.96
1	1	0	1 438	Seuil non chaînage	11	28.79	71.21
0	1	1	725		13.2	78.66	21.34
1	0	1	291	Seuil chaînage	15.2	96.68	3.32
1	1	1	8 852		24.4	99.99	4e-04

FIGURE 6 – Seuils selon le poids composé. $P(m)$: Probabilité que les 2 enregistrements de la paire correspondent au même individu. $G(u)$: Probabilité que les 2 enregistrements correspondent à 2 individus différents.

	Nom	Prénom	Date de naissance	
	`fe1fb20e56bd...`	`5b7808252fec...`	`aeed71d1dc67...`	
	`fe1fb20e56bd...`	`5b7808252fec...`	`9b1549d98eab...`	
Poids	+8.4	+5.7	-3.1	= 11

FIGURE 7 – Exemple de calcul de poids composé pour deux enregistrements anonymisés.

peut effectuer successivement plusieurs appariements par blocs selon plusieurs champs, ce afin de trouver davantage d'appariements : si le champ de sexe n'est pas fiable, on trouvera des appariements supplémentaires en procédant par bloc selon l'année de naissance, ou réciproquement.

3 Chiffrement

3.1 Principes

Les techniques de chiffrement[14] (*enciphering*) consistent à rendre un message illisible pour les personnes n'ayant pas la clé pour le rendre à nouveau lisible. Il s'agit d'un domaine de recherche très actif, car ces techniques sont à la base de la sécurité des communications sur internet, des transactions bancaires, etc. A la différence du hachage, le chiffrement est réversible.
On distingue deux familles de techniques de chiffrement : celles utilisant la même clé pour chiffrer et déchiffrer (méthodes dites symétriques) et celles utilisant deux clés, l'une publique et l'autre privée (méthodes dites asymétriques ou "à clé publique").

3.2 Méthodes symétriques

Avec les méthodes n'utilisant qu'une seule clé, l'émetteur et le récepteur du message chiffré doivent avoir la même clé, gardée secrète. Toutes les personnes ayant en leur possession

la clé et le message chiffré sont en mesure de déchiffrer le message et donc accéder à l'information qu'il contient.
Ces méthodes ont plusieurs inconvénients. Elles nécessitent que l'émetteur et le récepteur utilisent un canal sécurisé pour partager la clé. D'autre part, chaque paire ou groupe d'individus partageant des messages secrets doivent avoir la même clé réservée pour communiquer avec ce groupe d'individus seulement.
Ainsi, si Alice, Bob et Charlie veulent partager des messages deux à deux, il leur faut chacun deux clés. Pour chaque nouvelle personne avec qui ils souhaitent communiquer, il leur faut chacun une clé supplémentaire, et cela sans compter les combinaisons possibles de composition de groupes de personnes avec qui échanger. Il devient rapidement difficile de gérer toutes les clés.

3.3 Méthodes asymétriques

Les méthodes de chiffrement asymétriques permettent de résoudre ce problème. En effet, ces techniques se basent sur des paires de clés, l'une publique et l'autre privée. Chaque personne possède une telle paire de clés. La clé privée, comme son nom l'indique, n'est pas partagée et reste en la seule possession de son propriétaire. La clé publique, quant à elle, peut être associée au propriétaire dans le cadre d'un annuaire d'authentification, de façon à assurer que cette clé est bien la clé publique correspondant à la clé privée du propriétaire. Cependant, chacun peut avoir autant de paires de clés qu'il désire et diffuser la partie publique comme il le souhaite.
Lorsqu'un message est chiffré avec la clé privée, seule la clé publique permet de le déchiffrer. Cela permet de signer électroniquement un message pour authentifier son auteur [15].
Mais il est également possible de chiffrer un message avec la clé publique. Dans ce cas, seul le détenteur de la clé privée associée pourra déchiffrer le message, ce qui permet la *confidentialité* des échanges.
Dans la suite, nous utiliserons les notations suivantes :

- pub_X désignera la clé publique de X ou la partie publique d'une clé X,
- $priv_X$ désignera la clé privée de X ou la partie privée d'une clé X,
- $C_k(I)$ désignera le chiffrement de l'information I en utilisant la clé k ; si k est une clé privée, l'information sera chiffrée pour authentification ; si k est une clé publique, l'information sera chiffrée pour confidentialité ;
- $C_k^{-1}(I)$ désignera le déchiffrement de l'information chiffrée I en utilisant la clé k.

On aura donc les relations suivantes :

- $C^{-1}_{priv_X}\left(C_{pub_X}(I)\right) \rightarrow I$ (confidentialité),
- $C^{-1}_{pub_X}\left(C_{priv_X}(I)\right) \rightarrow I$ (authentification).

Il est possible de combiner authentification et confidentialité. Si Alice veut transmettre une information secrète à Bob, tout en permettant à Bob de s'assurer que cette information vient bien d'Alice, cette dernière utilisera sa clé privée à elle pour signer le message et la clé publique de Bob pour chiffrer le tout. A la réception, Bob utilisera sa clé privée

15. L'authentification peut se faire de la façon suivante. Lorsque Bob souhaite envoyer un message à Alice en permettant à cette dernière d'être sûre que c'est lui qui l'envoie, il applique une fonction de hachage à son message, pour obtenir une empreinte (ou condensat). Puis il chiffre cette empreinte avec sa clé privée. A la réception du message, Alice peut à son tour appliquer la même fonction de hachage sur le message, puis déchiffrer le condensat à l'aide de la clé publique de Bob, et enfin comparer les deux condensats. S'ils sont identiques, c'est bien Bob l'auteur du message, puisque seul le propriétaire de la clé privée (Bob) a pu le chiffrer de sorte que la clé publique (de Bob) puisse le déchiffrer. On peut donc *authentifier* l'auteur d'un message.

pour déchiffrer le message et la clé publique d'Alice pour s'assurer que le message est bien d'elle.
Une autre combinaison est également possible. Ainsi, si l'on souhaite qu'une information I ne soit accessible que lorsque deux personnes A et B donnent leur accord, il suffit de chiffrer cette information successivement avec deux clés publiques. L'accès à l'information initiale requiert alors l'utilisation, en ordre inverse, des deux clés privées (l'une en possession de A, l'autre en possession de B) correspondant aux clés publiques :

$$\begin{aligned} C^{-1}_{priv_A}\left(C^{-1}_{priv_B}\left(C_{pub_B}(C_{pub_A}(I))\right)\right) &\rightarrow C^{-1}_{priv_A}\left(C_{pub_A}(I)\right) \\ &\rightarrow I \end{aligned}$$

D'autres combinaisons sont possibles mais dans la suite c'est surtout la confidentialité permise par ce système de chiffrement qui nous intéresse.

3.4 Utilisation pour les appariements

Ces méthodes cryptographiques permettent de sécuriser les échanges de données, en assurant à la fois leur confidentialité et leur origine.
Elles ne permettent pas l'anonymisation de données, étant donné le caractère réversible du chiffrement. Cependant, combinées au hachage, elles peuvent permettre de confier l'appariement de données à un tiers de confiance, tout en cloisonnant l'accès aux données personnelles. C'est ce que nous exposons dans la section suivante.

4 Un système d'information statistique

La situation est donc la suivante : Les administrations sont riches d'informations médicales, scolaires, sociales, ..., dont certaines utilisent le NIR comme identifiant.
L'utilisation du NIR pour le croisement de fichiers doit faire l'objet d'un décret en Conseil d'Etat. Cependant, les techniques de hachage des champs identifiants permettent une anonymisation relative. L'utilisation de ces techniques permet le croisement de fichiers utilisant un NIR ainsi anonymisé sans requérir un décret du Conseil d'Etat[8, 24].
Reste qu'il ne suffit pas de pseudonymiser des champs identifiants (comme le NIR) pour garantir un certain niveau d'anonymat. En effet, si des données même pseudonymisées peuvent parfois encore permettre la ré-identification via notamment des informations de trajectoire, ce risque est encore plus grand lorsque d'autres informations sont agrégées suite à un appariement. On pourra lire à ce sujet [13] [16].
On souhaite donc avoir un tiers de confiance réalisant l'appariement et les études statistiques requises, et n'ayant accès qu'au minimum de données nécessaires pour l'appariement et l'étude en question, étude pour laquelle il aura eu l'autorisation de la CNIL.
Dans [18], une organisation est proposée dans ce sens.

4.1 Principe

Il s'agit de satisfaire deux contraintes. L'une est le partage d'identifiants communs pour permettre l'appariement. L'autre est la contrainte d'anonymat des données.

16. Signalons au passage l'existence du projet de recherche Cappris sur la protection des données privées, qui collabore déjà avec l'INSEE. https://cappris.inria.fr/

Dans cette section, nous prendrons l'exemple du NIR comme identifiant utilisé pour l'appariement. La section suivante discutera de la généralisation de cette technique à d'autres identifiants.
En appliquant un double hachage au NIR, ce dernier peut-être utilisé pour le croisement de fichiers après autorisation de la CNIL [8]. Il faut bien sûr que les clés de hachage utilisées pour hacher le NIR dans les deux fichiers à croiser soient les mêmes. Cela implique que si une entité se trouve en possession de deux fichiers avec des NIR hachés de la même façon, le risque de ré-identification augmente.
L'idée est donc d'utiliser les techniques de chiffrement sur le NIR doublement haché pour que seul un tiers autorisé puisse, en déchiffrant ce champ, revenir au NIR doublement haché dans chacun des fichiers et effectuer l'appariement d'après ce champ.
Pour [11], une telle approche aurait facilité l'appariement entre le fichier historique des demandeurs d'emploi de Pôle Emploi et les DADS [17] disponibles à l'INSEE.
Dans cette optique, la procédure d'appariement serait la suivante :

1. Pôle Emploi et l'INSEE effectuent respectivement chacun un double hachage du NIR (dénoté DH(NIR) dans la suite) dans le fichier des demandeurs d'emploi et dans celui des DADS,
2. L'organisme responsable de l'appariement, en l'occurence la Dares (ou une Autorité de gestion des clés, cf. section 4.3), crée pour cette étude une paire de clés dont elle conserve la clé privée ($priv_{Dares}$) ; la clé publique (pub_{Dares}) est envoyée à Pôle Emploi et l'INSEE,
3. Pour chaque enregistrement de leur fichier respectif, une clé I_k est créée à partir de l'identité de l'individu et utilisée pour chiffrer DH(NIR). Le couple $(I_k, C_{I_k}(DH(NIR)))$ est ensuite chiffré avec pub_{Dares} et l'enregistrement identifiant contient donc

$$C_{pub_{Dares}}(I_k, C_{I_k}(DH(NIR)))$$

4. Pôle Emploi et l'INSEE envoient les fichiers ainsi anonymisés et chiffrés à la Dares,
5. La Dares étant seule détentrice de la clé $priv_{Dares}$, elle seule peut, pour chaque enregistrement, effectuer les opérations suivantes nécessaires pour retrouver le $DH(NIR)$:

$$\begin{aligned} C^{-1}_{priv_{Dares}}\left(C_{pub_{Dares}}(I_k, C_{I_k}(DH(NIR)))\right) &\rightarrow (I_k, C_{I_k}(DH(NIR))) \\ C^{-1}_{I_k}(DH(NIR)) &\rightarrow DH(NIR) \end{aligned}$$

6. Il ne reste à la Dares qu'à apparier d'après le champ DH(NIR) de chaque enregistrement.

A la fin de cette procédure, la Dares n'a pas accès aux informations nominatives, puisque seul le NIR doublement haché lui est accessible. Si les clés utilisées pour le double hachage du NIR restent connues seulement de Pôle Emploi et l'INSEE, la Dares ne peut revenir au NIR, même par une attaque par dictionnaire.
De leur côté, ni Pôle Emploi ni l'INSEE ne peuvent avoir accès à des informations supplémentaires, car même s'ils entraient en possession du fichier transmis par l'autre organisme à la Dares, le double chiffrement les empêche de revenir au NIR doublement haché. En effet, chaque $DH(NIR)$ étant chiffré par une clé différente, le même $DH(NIR)$ sera chiffré différemment pour chaque enregistrement, ce qui prévient contre les attaques

17. Déclaration Annuelle de Données Sociales.

par dictionnaire. Sans ce chiffrement unique (dont la clé de déchiffrement I_k est elle-même chiffrée et accessible uniquement à la Dares), un NIR (déjà doublement haché de la même façon à Pôle Emploi et à l'INSEE) serait chiffré de la même façon par la clé pub_{Dares}, donc il serait possible d'apparier selon le résultat de ce chiffrement.
Cette procédure préserve donc bien l'anonymat des données[18] tout en permettant un appariement relativement aisé.

4.2 Application à des fichiers sans identifiant unique commun

La méthode exposée dans la section précédente peut être généralisée à n'importe quelle information identifiante à la place du NIR. Si les fichiers à apparier ne contiennent pas d'information commune identifiant uniquement un individu, il reste possible de réaliser l'appariement sur plusieurs champs comme le nom, le prénom, la date et la commune de naissance, ... Sur chacun des champs sont alors appliqués un double hachage et un double chiffrement. Le tiers réalisant l'appariement peut ensuite, après déchiffrement de chaque champ, appliquer une méthode d'appariement probabiliste (cf. section 2.2), dont on a vu qu'elle restait efficace même sur des données anonymisées[15].

4.3 Gestion des clés

Cette solution nécessite une gestion rigoureuse des clés utilisées pour le hachage et le chiffrement, pour plusieurs raisons.
La première est la possibilité de mener des études sur de longues périodes. Pour garantir une disponibilité des données des administrations sur le long terme, elles pourraient être régulièrement archivées en les anonymisant, notamment par un hachage du NIR et autre données identifiantes, et en les chiffrant. Ainsi, si des études nécessitent de croiser des données récentes avec des données archivées, il suffit alors de faire subir aux données récentes le même hachage que les données archivées et déchiffrées, d'où la nécessité de conserver les clés utilisées par le passé pour le hachage et le chiffrement.
La seconde raison tient au besoin de pouvoir reproduire des expériences et des études ou d'élargir le champ d'une étude par un appariement initialement non prévu, puis autorisé par la CNIL[19]. Il s'agit donc de garder une trace des études et des clés utilisées pour les mener.
Distinguons donc les deux hachages successifs à appliquer aux données. Le premier doit utiliser une clé commune à toutes les institutions, permettant notamment l'archivage sans empêcher l'utilisation de ces données archivées par la suite. Les données hachées en utilisant cette première clé commune ne fournissent pas un niveau d'anonymisation suffisant puisque toutes les institutions disposent de la clé, rendant possible les attaques par dictionnaire (cf. section 1.2). Cependant, le fait de les chiffrer permet de les protéger, pour peu que la clé permettant leur déchiffrement soit conservée en sécurité.
Le second hachage doit être spécifique à chaque étude, afin de garantir un niveau suffisant d'anonymisation par pseudonymisation : ainsi, en gardant secrète la seconde clé de ha-

18. Anonymat toujours relatif, rappelons-le, selon les autres données présentes et qui permettent éventueller de reconstituer des trajectoires uniques et donc identifiantes si un tiers a de son côté assez d'informations pour les recouper.

19. Par exemple, on peut imaginer qu'une étude sur le diabète pourrait être reprise par la suite par des chercheurs en ophtalmologie pour être enrichie en utilisant des données supplémentaires. Cela ne sera possible que si les identifiants des nouvelles données peuvent subir le même traitement de pseudonymisation que celui utilisé pour l'anonymisation des données de l'étude originale.

chage, les croisements faits dans une étude ne sont utilisables que pour cette étude. Si on souhaite reprendre cette étude pour l'enrichir avec de nouvelles données, il serait possible de le faire en obtenant d'une part l'autorisation de la CNIL, d'autre part la clé du second hachage utilisée pour l'étude. A ce moment-là, un premier hachage par la clé commune à toutes les institutions puis un second hachage par la clé spécifique à l'étude permet de croiser les données complémentaires avec celles de l'étude.
Enfin, il convient de s'assurer que les clés utilisées sont authentiques : Lorsqu'un établissement communique à un autre des données chiffrées à l'aide d'une clé publique, il doit être sûr que cette clé est bien celle à utiliser pour transmettre les données à l'établissement de destination et pour l'étude en question.
Une Autorité de gestion des clés est donc nécessaire. Une telle Autorité peut se voir confier les clés de hachage utilisées par un ou plusieurs organismes, associées aux études menées. Cette Autorité peut également générer des clés de hachage ou de chiffrement. Dans ce dernier cas, pour la procédure exposée dans la section 4.1, elle peut envoyer la clé publique aux fournisseurs de données (pour qu'ils chiffrent les données avec cette clé) et la clé privée (correspondant à la clé publique) au seul organisme habilité à déchiffrer les données pour appariement. Enfin, cette autorité peut conserver les clés permettant de déchiffrement des données archivées.
Dans le cas où l'Autorité de gestion des clés fournit les clés, elle les signe pour les authentifier et les chiffre pour assurer leur confidentialité et leur intégrité lors de leur transmission aux établissements destinatires.
La chaîne de manipulation des données est alors sécurisée par le chiffrement tandis que l'anonymat des données est respecté, puisque les données identifiantes sont doublement hachées avant transmission au tiers réalisant l'appariement.
Avec une telle Autorité de gestion des clés, la procédure décrite en exemple dans la section 4.1 devient la suivante :

1. L'Autorité de gestion des clés envoie à Pôle Emploi et à l'INSEE la clé spécifique à l'étude pour le second hachage du NIR ainsi que la partie publique de la clé de chiffrement de l'étude (Pub_{Dares}) ; elle envoie également la partie privée de cette clé de chiffrement ($Priv_{Dares}$) à la Dares (figure 8),
2. Pour chaque enregistrement de leur fichier respectif, Pôle Emploi et l'INSEE applique le double hachage du NIR (d'abord avec la clé commune à toutes les institutions, puis avec la clé spécifique à l'étude), puis le chiffrement du DH(NIR) avec une clé nouvelle pour chaque enregistrement, puis le chiffrement, par Pub_{Dares}, de cette clé et du DH(NIR) chiffré (figure 9). Le résultat nécessite la clé $Priv_{Dares}$ pour être déchiffré, dont seul la Dares [20] est en possession,
3. Pôle Emploi et l'INSEE envoient leurs fichiers ainsi hachés et chiffrés à la Dares (figure 10),
4. La Dares procède au double déchiffrement de chaque DH(NIR) pour effectuer les appariements. Elle ne peut cependant remonter au NIR, à cause du double hachage (figure 11).

20. L'Autorité de gestion des clés a également la clé $Priv_{Dares}$ de cette étude, mais elle n'a bien sûr par les données chiffrées.

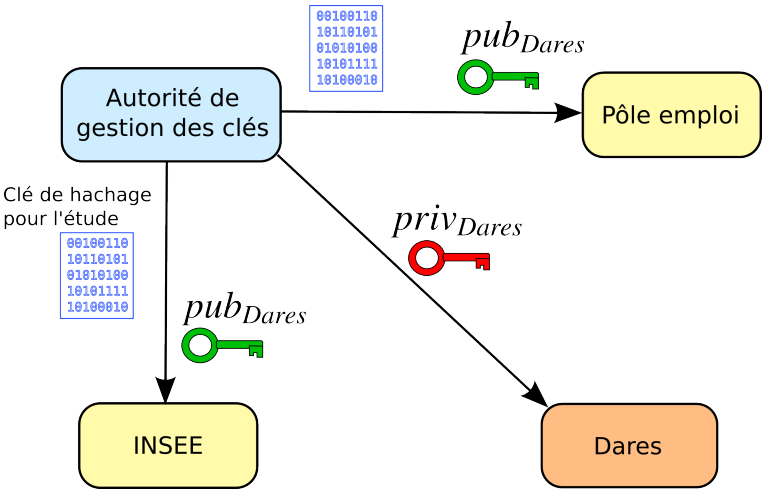

FIGURE 8 – Distribution des clés par l'Autorité de gestion des clés.

5 Conclusion

Comme nous l'avons vu, les techniques informatiques actuelles (hachage et chiffrement) permettraient, tout en conservant un niveau d'anonymat suffisant pour satisfaire les exigences de la CNIL, de réaliser des études statistiques nécessitant des appariements de fichiers sociaux et médicaux.

La proposition exposée plus haut, utilisant ces techniques, permettrait de mener plus facilement de telles études, plus régulièrement et de manière plus fine, tout en conservant un niveau d'anonymat suffisant. Elle suggère que les fichiers sociaux ou médicaux soient transmis à l'organisme réalisant l'appariement suite au double hachage du NIR : un hachage avec une clé commune à toutes les institutions suivi d'un hachage avec une clé spécifique à chaque étude, le tout chiffré avec une clé spécifique à chaque étude, les clés utilisées pour le hachage et le chiffrement étant conservées dans la confidentialité par une Autorité de gestion des clés.

Il nous semble important qu'une organisation du type de celle proposée soit mise en place pour débloquer les recherches et études utilisant des données sociales et médicales. L'obstacle à cette mise en place ne nous paraît pas d'ordre technique mais plutôt d'ordre organisationnel, dans la mesure où cette solution repose sur l'existence d'une Autorité de gestion des clés, dont le rôle est de générer, transmettre et conserver les clés de chaque étude. Cela pose la question de la création d'un organisme supplémentaire ou de confier le rôle de cette Autorité à un organisme existant (CNIS, CNIL, ...).

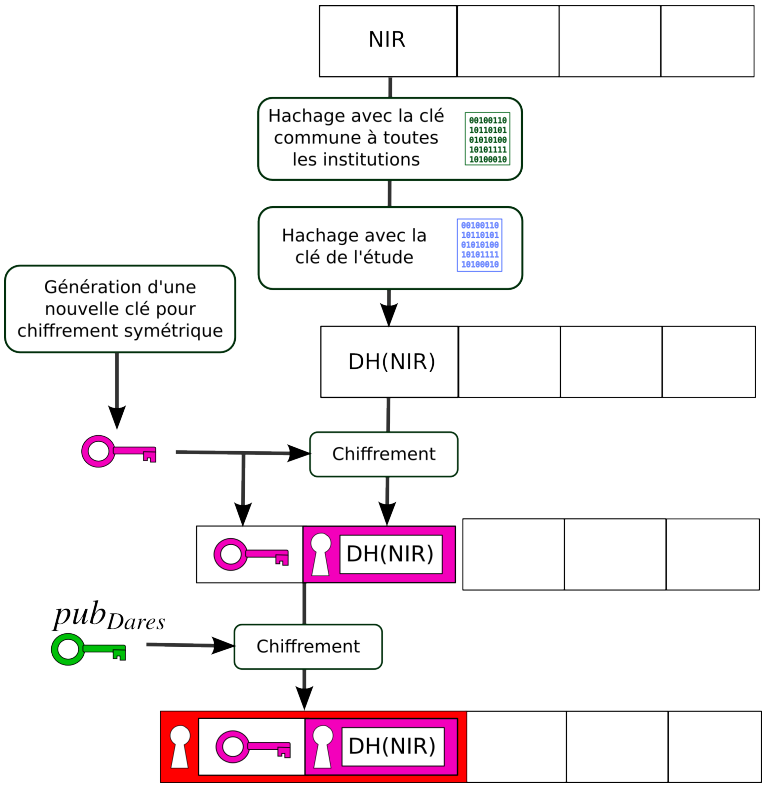

FIGURE 9 – Hachage et chiffrement du NIR de chaque enregistrement par Pôle emploi et l'INSEE.

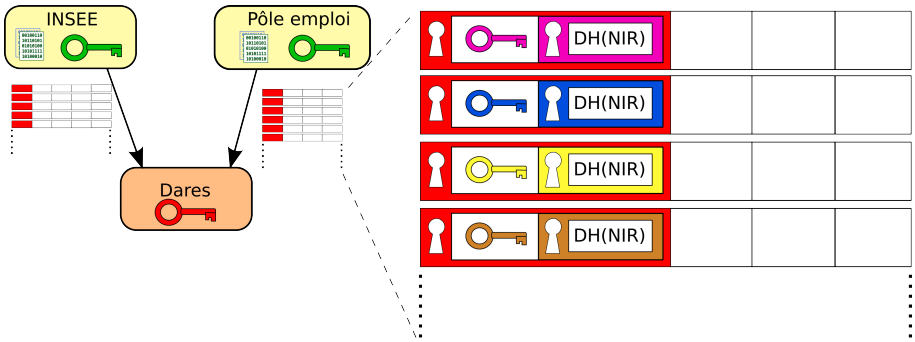

FIGURE 10 – Transmission sécurisée des données anonymisées.

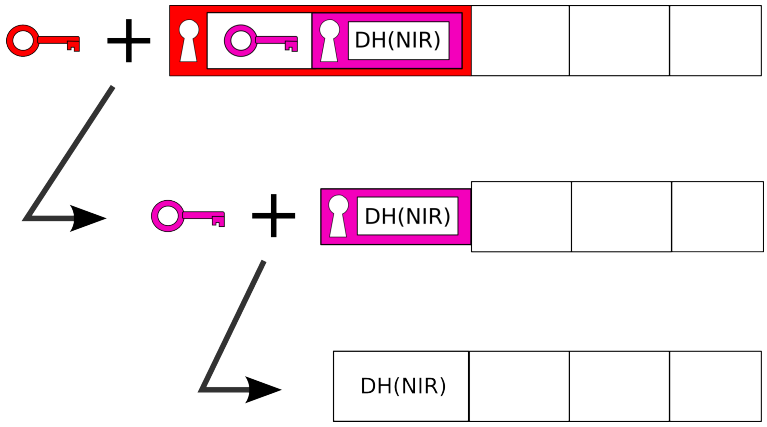

FIGURE 11 – Déchiffrements successifs à la Dares, en utilisant $C_{priv_{Dares}}$ sur chaque champ identifiant chiffré, puis en utilisant la clé dans ce champ pour déchiffrer le NIR doublement haché DH(NIR).

Références

[1] Agence nationale de la sécurité des systèmes d'information, *Référentiel Général de Sécurité version 2.0*, Version 2.03 du 21 février 2014, http://www.ssi.gouv.fr/IMG/pdf/RGS_v-2-0_B1.pdf

[2] Avarro, Gonzalo (2001), *A guided tour to approximate string matching*, ACM Computing Surveys 33 (1) : 31–88. doi :10.1145/375360.375365.

[3] Bounebache Said Karim, Rey Grégoire, Quantin Catherine, Riandey Benoit (en préparation), *Une revue des méthodes d'appariement : Applications et perspectives dans le cas des données de Santé*, En préparation.

[4] Cappé Olivier, Moulines Eric (2009), *On-line Expectation-Maximization Algorithm for Latent Data Models*, Journal of the Royal Statistical Society Series B (Statistical Methodology) 71, 3 (2009) 593-613. http://arxiv.org/abs/0712.4273

[5] Fellegi I.P., Sunter A.B. (1968), *A Theory for Record Linkage*, J. Am. Statistical Assoc., Vol. 64, no. 328 , 1183-1210.

[6] Fournel I., Schwarzinger M., Binquet C., Benzenine E., Hill C., Quantin C. (2009), *Contribution of Record Linkage to Vital Status Determination in Cancer Patients*, Studies in health technologies and informatics 150 :91-5

[7] Fox Karla, Stratychuk Lori (2010), *Méthodes de couplage d'enregistrements*, atelier du Symposium 2010 de Statistique Canada (26 octobre).

[8] Gensbittel Michel-Henri, Riandey Benoît, Quantin Catherine (2007), *Appariements sécurisés : statisticiens, ayez de l'audace !*, Courrier des statistiques 121-122, http://www.insee.fr/fr/themes/document.asp?reg_id=0&id=2153

[9] Guillerm Marie (2009), *L'appariement de fichiers pour le suivi de l'étudiant*, DEPP – Bureau des études statistiques sur l'enseignement supérieur

[10] Jaro Matthew A. (1995),*Probabilistic linkage of large public health data files*, Statistics in Medicine, 14 : 491–498, doi : 10.1002/sim.4780140510 DATA FILES

[11] Le Barbanchon Thomas, Sédillot Béatrice (2011), *L'appariement expérimental entre le fichier historique des demandeurs d'emploi et les DADS : premier bilan et perspectives*, Courrier des statistiques 131, http://www.insee.fr/fr/themes/document.asp?reg_id=0&id=3388

[12] National Institute of Standards and Technology, *Secure Hash Standard (SHS)*, Federel Information Processing Standards Publication, 2012, http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf

[13] Nguyen Benjamin (2014), *Techniques d'anonymisation*, Statistiques et société, Vol. 2, numéro 4, http://publications-sfds.fr/index.php/stat_soc/article/view/398

[14] Paar Christof, Pelzl Jan (2010) *Understanding cryptography, A Textbook for Students and Practitioners*, Springer, ISBN :978-3-642-04100-6

[15] Quantin C., Bouzelat H., Allaert F.-A., Benhamiche A.-M., Faivre J., Dusserre L. (1998), *Automatic record hash coding and linkage for epidemiological follow-up data confidentiality*, Methods of information in medicine 37(3) :271-7.

[16] Quantin Catherine, Gouyon Béatrice, Allaert François-André, Cohen Olivier (2005), *Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat : application au suivi des informations médicales*, Courrier des statistiques 113-114, http://www.insee.fr/fr/themes/document.asp?id=1501

[17] Quantin C., Binquet C., Allaert F.-A., Cornet B., Pattisina R., Le Teuff G., Ferdynus C., Gouyon J.-B. (2005), *Decision analysis for the assessment of a record linkage procedure : application to a perinatal network*, Methods of Information in Medicine, 44(1) :72-9.

[18] Quantin C., Fassa M., Coatrieux G., Trouessin G., Allaert F.-A. (2008), *Combining hashing and enciphering algorithms for epidemiological analysis of gathered data*, Methods of information in medicine, Vol. 47 Issue 5, pp 454-458, http://dx.doi.org/10.3414/ME0546

[19] Quantin C, Jaquet-Chiffelle D.-O., Coatrieux G., Benzenine E., Allaert F.-A. (2011), *Medical record search engines, using pseudonymised patient identity : an alternative to centralised medical records*, International Journal of Medical Informatics, 80 :e6-e11.

[20] Quantin C., Jacquet-Chiffelle D.-O., Coatrieux G., Benzenine E., Auverlot B., Allaert F.-A. (2011), *Medical record : systematic centralization versus secure on demand aggregation*, BMC Medical Informatics and Decision Making, 11 :18

[21] Quantin C., Benzenine E., Allaert F.-A., Guesdon M., Gouyon J.-B., Riandey B. (2014), *Epidemiological and Statistical Secured Matching in France*, Statistical Journal of the IAOS 30, 255–261.

[22] Schnell Rainer, Bachteler Tobias, Reiher Jörg (2009), *Privacy-preserving record linkage using Bloom filters*, BMC Medical Informatics and Decision Making, http://dx.doi.org/10.1186/1472-6947-9-41

[23] Stinson Douglas R. (2005), *Cryptography : Theory and Practice, Third Edition*, Chapman and Hall/CRC, ISBN :9781584885085

[24] Vuillet-Tavernier Sophie (2000), *Réflexion autour de l'anonymat dans le traitement des données de santé*, Med. Droit 2000 ;40 :1–4.

[25] Winkler William E. (1988), *Using the EM Algorithm for Weight Computation in the Felligi-Sunter Model of Record Linkage*, Bureau of the Census, Washington,D.C. 20233.