Appariement de données pseudonymisées

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

[1]Département d'Information Médicale - CHU Dijon
[2]INRIA - Paris Saclay
[3]INSERM, CIC 1432, Dijon

Journées de Méthodologie Statistique de l'Insee, Avril 2015

Plan

Introduction

- 1974 : Scandale SAFARI $\Rightarrow$ 1978 : Loi Informatique et Libertés, CNIL,
- NIR déjà utilisé dans les secteurs du travail, de la santé et les administrations sociales,
- Pour les autres, création d'identifiants ≪sectoriels≫ (éducation, finances, . . .),
- Le croisement de fichiers en utilisant le NIR nécessite un décret en conseil d'Etat. . .mais pas si le NIR est haché.

Anonymat

- Le traitement de données personnelles de santé à des fins d'études statistiques et de recherches nécessite en général qu'elles soient anonymisées,
- Pour le croisement de fichiers, anonymisation $\approx$ pseudonymisation,
- On a plutôt affaire à un ≪niveau d'anonymisation≫ :
  - $+$ on aggrège de données, $+$ le risque de réidentification augmente,
  - on ne connaît pas les données connues par un tiers.
- Les techniques informatiques actuelles permettant d'assurer un bon niveau d'anonymisation et de sécurisation des appariements.

Nous présentons la proposition de Quantin (2008) en épidémiologie et proposons de l'étendre à la statistique publique.

Plan

Hachage

Hachage = calcul d'une *empreinte* (ou signature) de taille fixe à partir de données de n'importe quelle taille.

La *distance* entre deux empreintes de deux données est indépendante de la distance initiale entre ces deux données :

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

Le risque de collision est faible (quasi nul).

Le hachage est **irréversible** : on ne peut retrouver x d'après hash(x), sauf par attaques dites ≪par dictionnaire≫.

Hachage - Résistance aux attaques

Attaque par dictionnaire : hachage de chaînes pour établir des correspondances chaîne $\Leftrightarrow$ empreinte. A partir d'une empreinte, on peut retrouver une chaîne originale possible. Le risque de collision étant faible, la chaîne correspondante est quasi-sûrement la chaîne originale.

Pour se prémunir : utilisation d'un **sel**, une chaîne secrète ajoutée à la donnée avant de la hacher.

Exemple : Avec comme sel "`XZ!#45`", on hachera non plus "`Dupont`" mais "`DupontXZ!#45`".

Le sel peut aussi être calculé à partir de la chaîne à hacher, selon une fonction secrète ou utilisant une clé secrète (comme dans le logiciel Anonymat).

L'important est que la procédure reste déterministe : la même entrée produit la même empreinte.

Hachage - Utilisation dans les appariements

Inconvénients :

- la moindre différence (erreur de saisie sur un nom, ...) donne deux empreintes radicalement différentes $\Rightarrow$ nécessite de normaliser les entrées,
- impossibilité d'utiliser des fonctions de distance entre deux identifiants.

Le double hachage, utilisant deux clés secrètes, offre un bon niveau de sécurité.

Plan

Appariement déterministe

1. Déterminer les champs identifiants dans les deux sources de données à apparier,
2. Définir une mesure de distance et un seuil à partir duquel deux enregistrements sont considérés comme correspondant au même individu.

L'utilisation la plus simple consiste à n'apparier que lorsque la distance entre deux identifiants est nulle (identifiants strictements identiques), par exemple pour un appariement par un NIR haché.

Appariement probabiliste

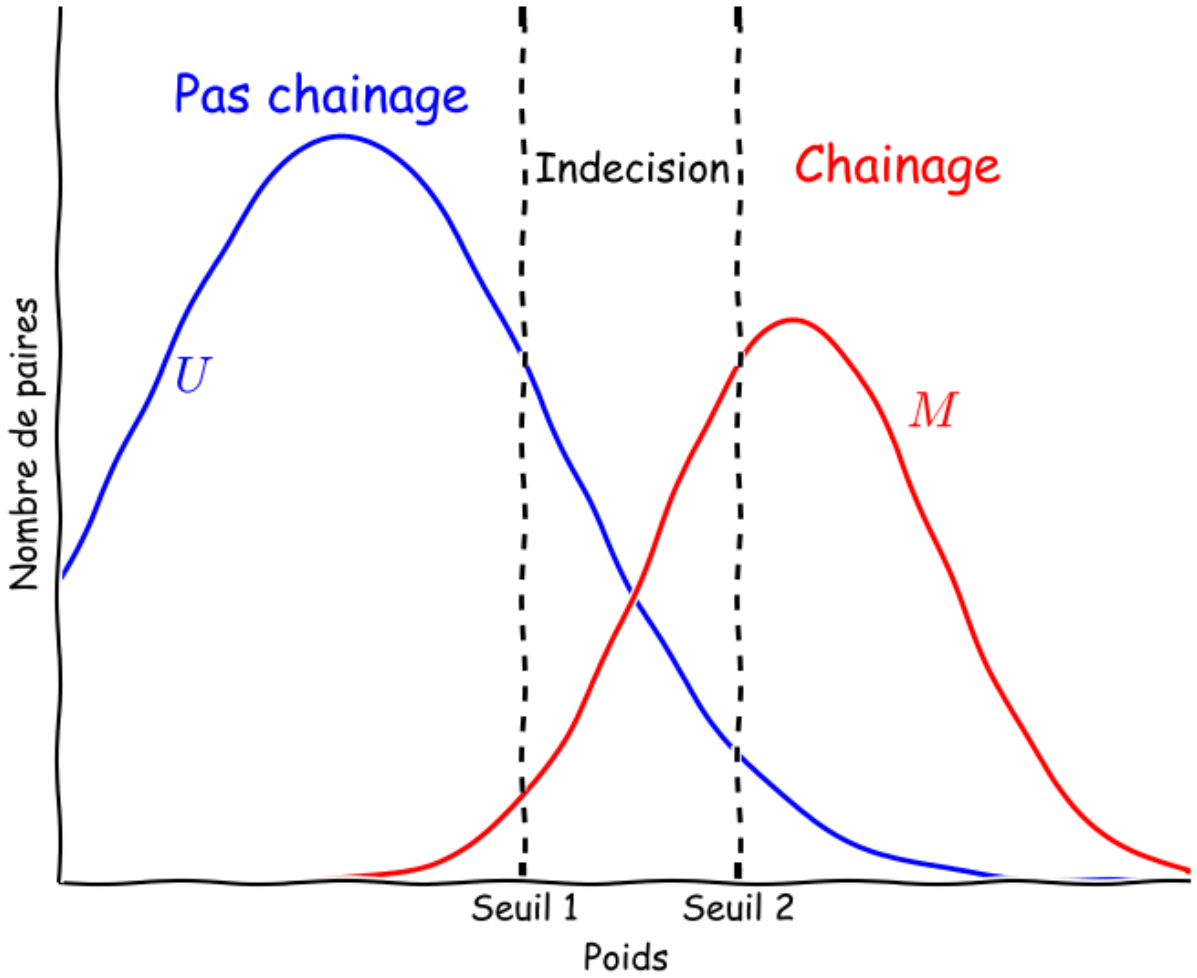
Utile lorsqu'on ne dispose pas d'un champ identifiant sans ambiguité les individus (comme le NIR) qui soit commun aux deux sources de données à apparier, et que l'établissement de règles (par exemples à partir de distances entre champs) n'est pas possible, comme c'est notamment le cas lorsque les informations ont été anonymisées par hachage.

Le cadre théorique de ces méthodes d'appariement probabiliste a été posé par Fellegi et Sunter en 1968. En 1995, Jaro applique pour la première fois ces méthodes sur des données de santé à l'aide d'un programme informatique. En 1998, Quantin décrit la première application de la méthode de Jaro à des fichiers ayant été préalablement anonymisés par hachage.

Appariement probabiliste - Principes

- Modèle de mélange de distributions utilisant la variabilité et la fréquence des valeurs de chaque champ identifiant ...
- ... pour estimer deux poids unitaires pour chaque champ (l'un en cas d'égalité, l'autre en cas de différence entre deux enregistrements), estimation réalisée par l'algorithme EM,
- Comparaison de chaque paire d'enregistrements pour calculer un poids composé,
- Selon le poids composé et des seuils choisis pour l'étude, classification automatique en "Non chaînage", "Chaînage" ou "Indécision",
- Traitements supplémentaires pour chaîner ou non des cas en indécision et/ou valider les chaînages obtenus.

Appariement probabiliste - Principes (2)



Appariement probabiliste - Exemple

	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

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

Revue des méthodes d'appariements

On pourra consulter

≪ Une revue des méthodes d'appariement :
Applications et perspectives dans le cas des données de Santé ≫

(Bounebache Said Karim, Rey Grégoire, Quantin Catherine, Riandey Benoit – en préparation)

Plan

Chiffrement

Les techniques de chiffrement (*enciphering*) consistent à rendre un message illisible pour les personnes n'ayant pas la clé pour le rendre à nouveau lisible.

Deux familles :

- méthodes symétriques : même clé pour chiffrer et déchiffrer ; nécessite une clé par groupe en communication + problème de transmission confidentielle des clés ;
- **méthodes asymétriques ou ≪ à clé publique ≫** : utilisation de paires (clé publique, clé privée) ; ce qui est chiffré par une clé nécessite l'autre clé pour être déchiffré. Permet la confidentialité et l'authentification.

Chiffrement et anonymat

Les méthodes de chiffrement ne permettent pas l'anonymisation de données, puisqu'elles ne sont pas irréversibles : une identité chiffrée peut être déchiffrée à l'aide de la bonne clé.

Cependant, les méthodes de chiffrement asymétriques peuvent être utilisées pour sécuriser un processus d'appariement de données anonymisées (pseudonymisées).

Plan

Système d'information statistique

En 2008, une méthode utilisant hachage et chiffrement est proposée par Quantin *et al* pour effectuer des appariements dans les études épidémiologiques, garantissant un certain niveau d'anonymat et permettant un enrichissement ultérieur des études.

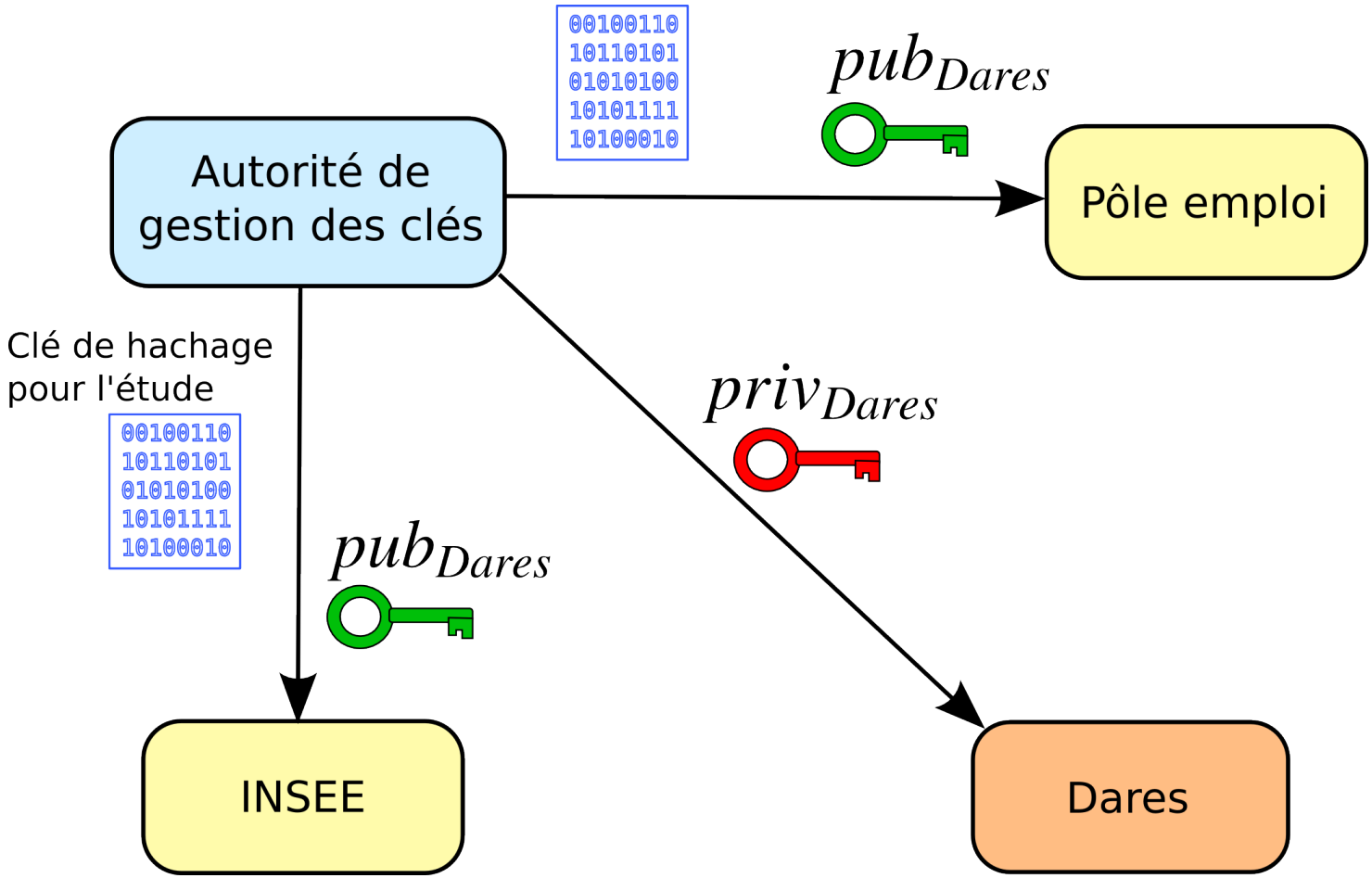
Nous proposons de l'étendre à la statistique publique.

Présentation sur un exemple inspiré de ≪ L'appariement expérimental entre le fichier historique des demandeurs d'emploi et les DADS : premier bilan et perspectives ≫ (Le Barbanchon, Sédillot).

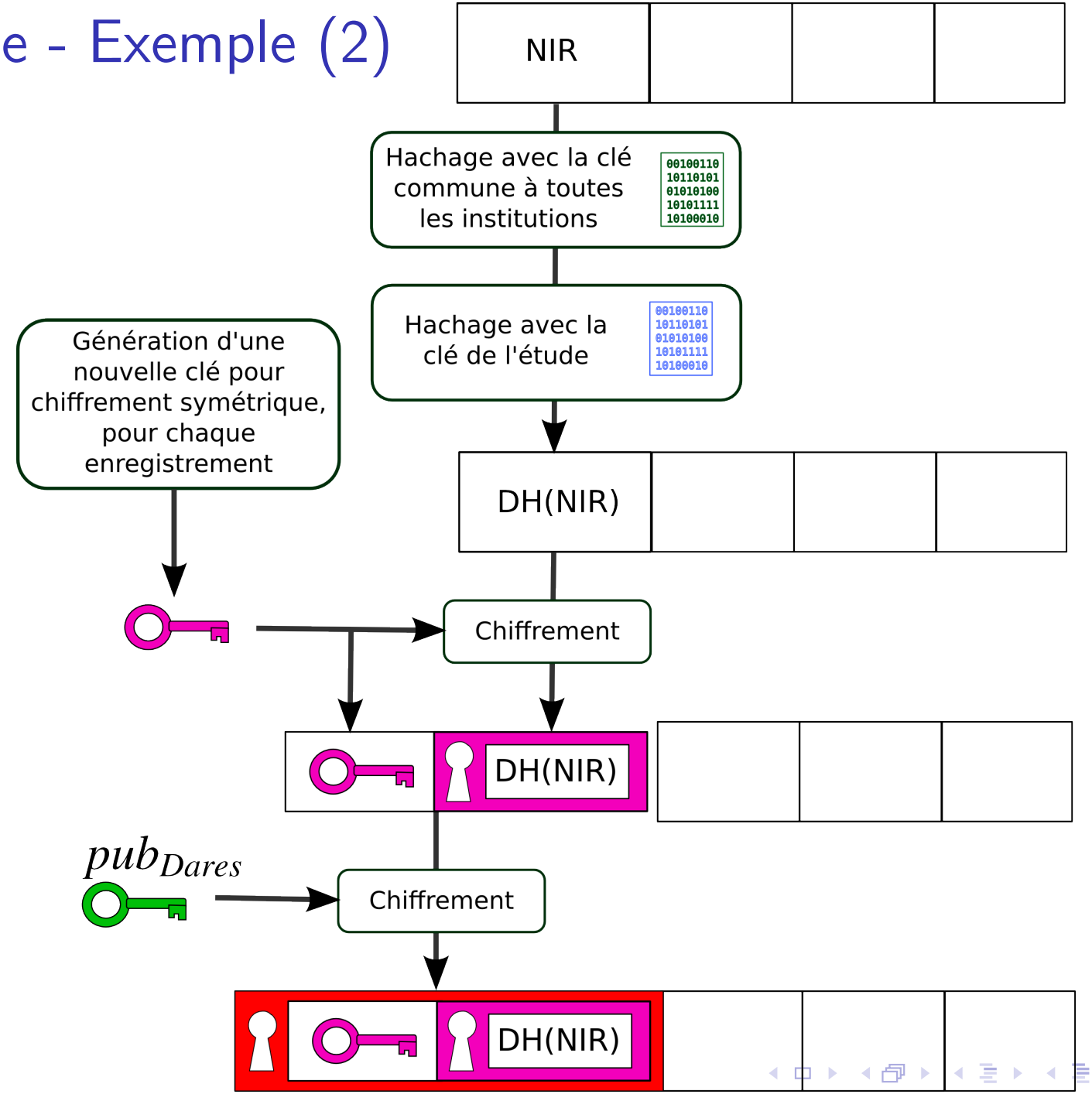
Méthode - Principes

- Chaque producteur de données :
  - Double hachage du NIR : hachage avec clé commune à toutes les institutions + hachage avec une clé secrète par étude ($\Rightarrow$ DH(NIR)),
  - Chiffrement de chaque DH(NIR) par une clé unique,
  - Chiffrement de cette clé + DH(NIR) chiffré par une clé publique unique à l'étude,
- Déchiffrement dans un organisme tiers, possédant la clé privée de l'étude,
- Une Autorité de gestion des clés génère, délivre et conserve les clés utilisées.

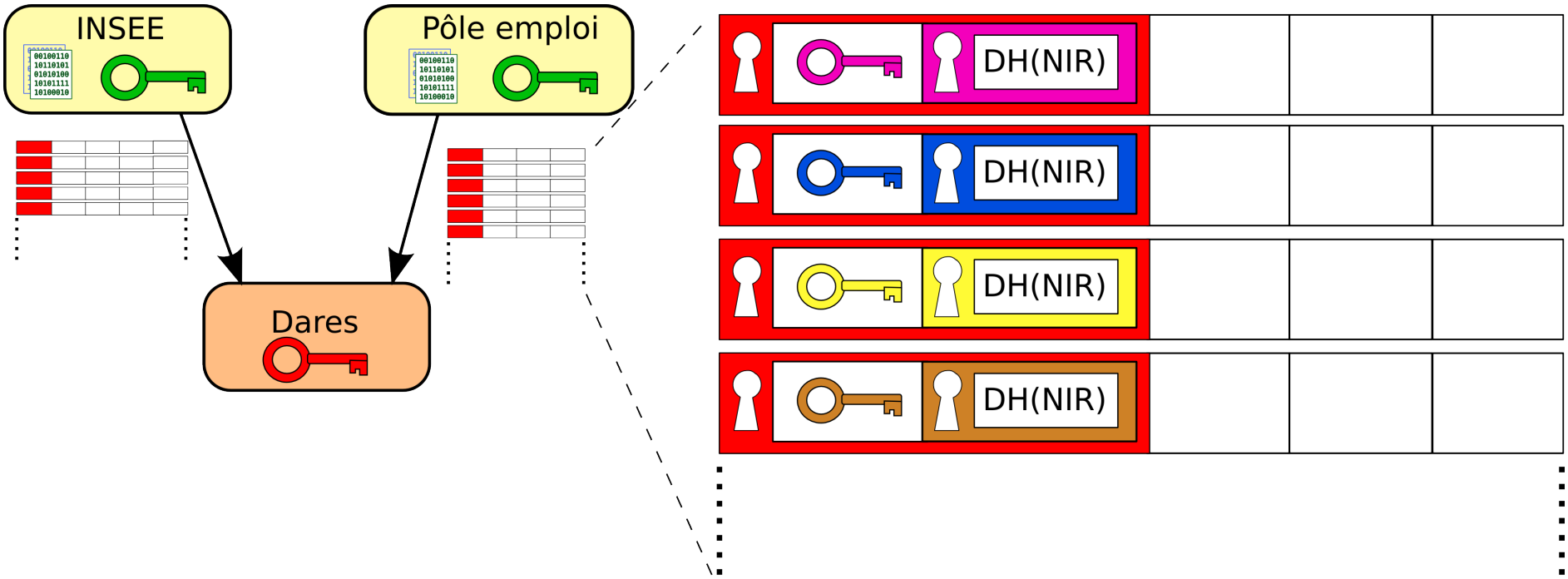
Méthode - Exemple



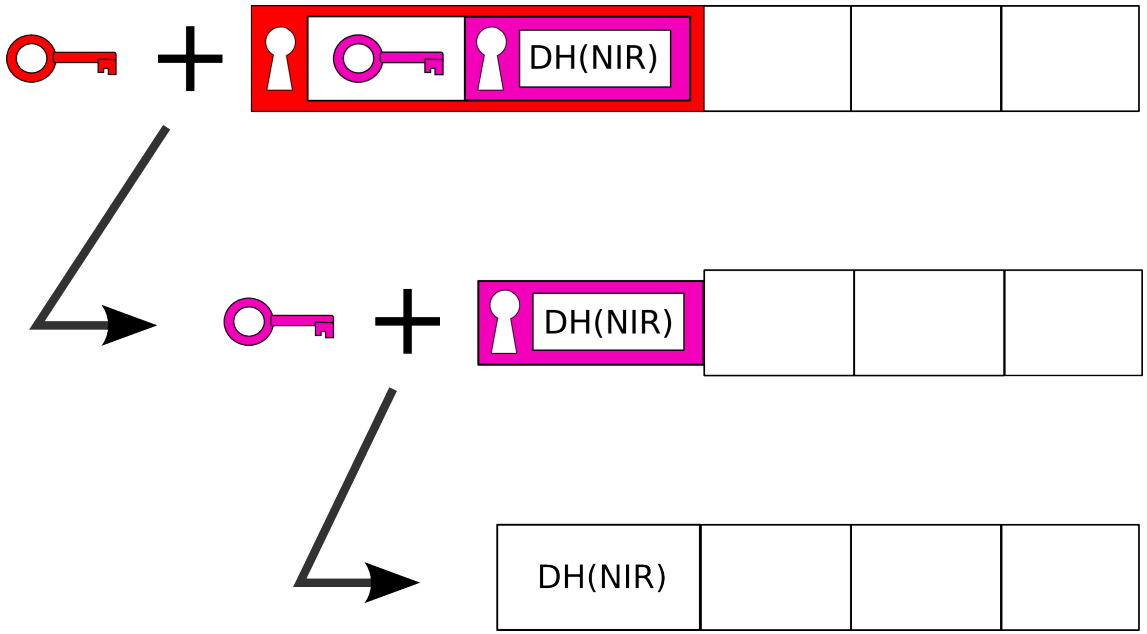
Méthode - Exemple (2)



Méthode - Exemple (3)



Méthode - Exemple (4)



Méthode - Remarques

- Les données restent anonymes : l'organisme croisant les données ne peut remonter aux informations identifiantes, doublement hachées,
- La méthode peut être utilisée pour des appariements probabilistes, en appliquant hachages et chiffrements sur chaque champ identifiant ; cela n'empêche pas le fonctionnement de la méthode de Jaro.
- La conservation des clés par l'Autorité de gestion des clés permet d'enrichir des études antérieures par de nouvelles données, en leur faisant subir le même hachage que l'étude d'origine.
- L'archivage des données administratives est fait après un hachage des identifiants avec la clé de hachage commune, suivi d'un chiffrement, dont la clé de déchiffrement est détenue par l'Autorité de gestion des clés.

Plan

Conclusion

- Hachage et chiffrement permettent la mise en place de procédures sécurisées d'appariement respectant l'anonymat...
- ... tout en n'obérant pas la possibilité d'enrichissement des études par de nouvelles données,
- La mise en place d'une telle solution permettrait de plus facilement réaliser des études de statistique publique et des recherches utilisant des données sociales et de santé,
- L'obstacle n'est pas technique mais organisationnel,
- La méthode implique la mise en place d'une Autorité de gestion des clés,
- Ce rôle pourrait être confié à un organisme existant (CNIS, CNIL, ...?).

Merci