Appariement de données médico-sociales: techniques et organisations

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

Séminaire Aristote - 15 octobre 2015

Problématique générale

Besoin d'appariements de données sociales et médicales pour

- Recherches et études statistiques,
- Recherches épidémiologiques,
- Recherches d'antécédents,
- ....

mais

- Multiples sources de données, nationales et locales : PMSI, Registres, ...,
- Pas toujours d'identifiant national utilisable,
- Respect de la loi Informatique et Libertés (et transposition européenne).

Anonymisation ?

- Rendre impossible l'identification d'une personne,
- Pseudonymisation : aléatoire ou déterministe, doit être irréversible,
- Assez d'information non directement identifiante peut permettre d'identifier (ex : parcours de soins) ; le ≪niveau d'anonymat≫ dépend des données dont on dispose,
- L'appariement de données nécessite que les données ne soient pas ≪complètement anonymes≫ : on doit pouvoir décider que deux jeux de données concernent le même individu,
- L'appariement de données augmente le risque de réindentification,
- Techniques : hachage, chiffrement, agrégation de données, dégradation des données, ...

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

Hachage - Utilisation dans les appariements

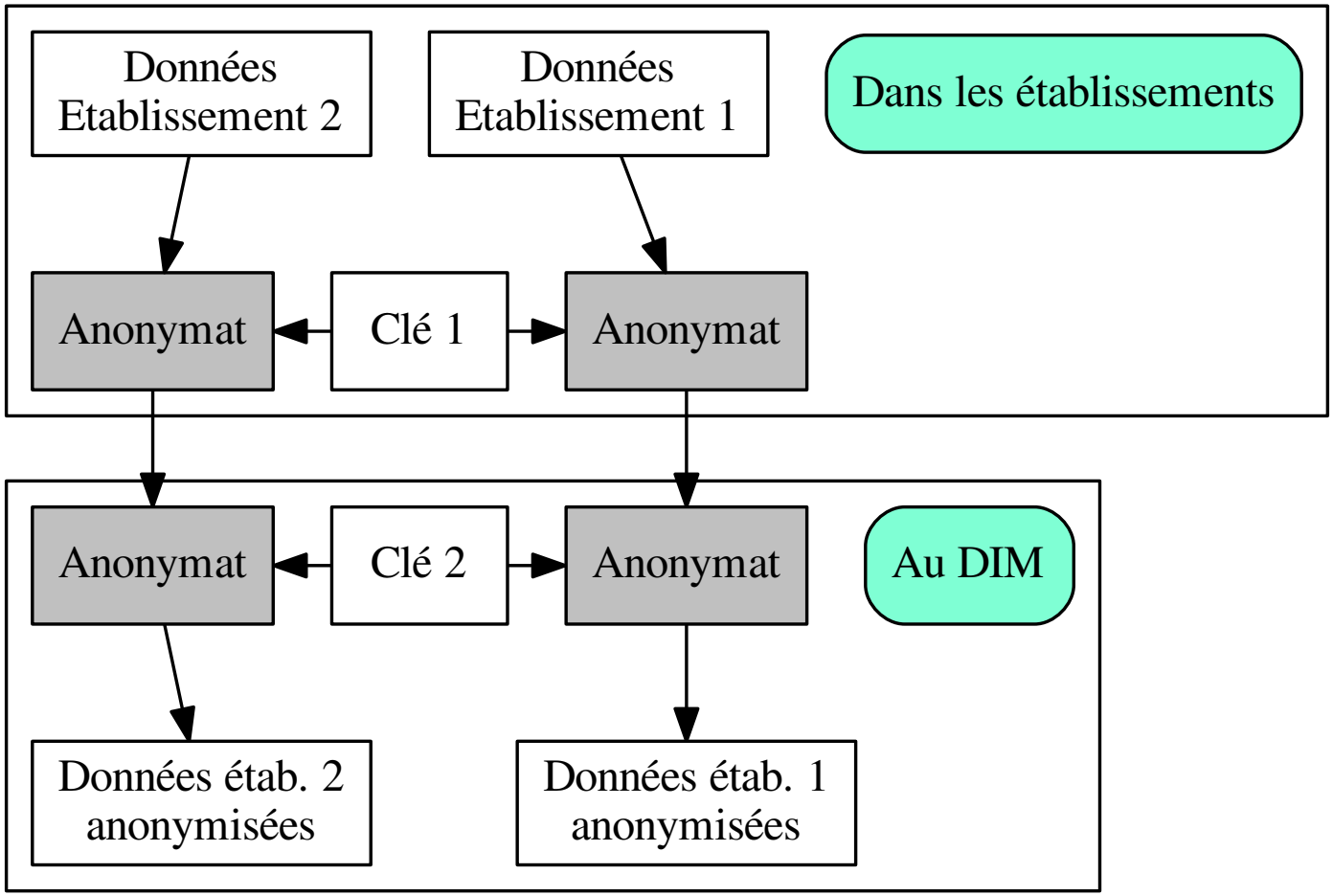
Inconvénients :

- 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.

Anonymisation des données - Exemple

Hachage de champs identifiants (SHA-1 ou SHA-256) avec clé secrête.
Déterminisme : x = y ⇒ Anonymat(x, clé) = Anonymat(y, clé)



Objectifs du chaînage (appariement)

Rapprocher les données d'un même individu 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.

Chaînage d'enregistrements composés de plusieurs variables (champs).
Types de chaînage :

- déterministe, quand les deux sources partagent un identifiant commun,
- probabiliste sinon, par exemple la méthode de Jaro.

Méthode probabiliste de Jaro

- Association d'un poids aux différents champs identifiants utilisés pour le chaînage,
- Ce poids diffère selon le pouvoir discriminant du champ ; par exemple, le sexe est moins discriminant que la date de naissance,
- Selon des seuils, traitement automatique aboutissant à chaînage, non chaînage ou zone d'indécision,
- Champs hachés $\Rightarrow$ pas d'utilisation de distances (on peut découper par groupes de lettres, mais plus long).

Calcul de la pondération pour chaque variable

Par exemple le nom :

	même individu	individus $\neq$
même nom	VP	FP = Collision
noms $\neq$	FN = Doublon	VN

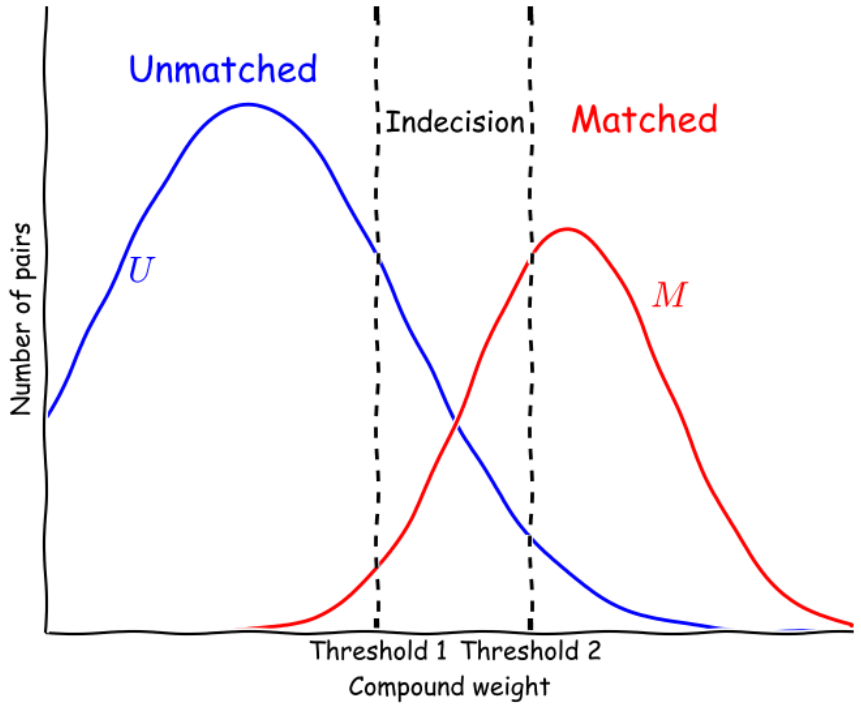
N paires d'enregistrements.
Rapport de vraisemblance $= \frac{1-\text{taux de doublons}}{\text{taux de collisions}}$

Modélisation statistique

2 sous populations théoriques :

- ensemble M des paires correspondant :
  - au même individu,
  - en proportion p,
- ensemble U des paires correspondant à des individus différents.

Modèle par mélange des 2 distributions pour estimer les taux de doublons et de collisions.



Exemple de calcul des taux

	Doublons (%)	Collisions (%)
Nom	6.053	0.021
Prénom	3.081	0.335
Date de naissance	4.469	0.003

Exemple de calcul des poids

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

- Poids attribué à chaque paire d'enregistrements en fonction
  - de la concordance des différents champs (nom, prénom, ...),
  - de la quantité d'information apportée par chaque champ.

  La concordance peut être calculée selon une méthode adaptée à chaque champ (par exemple comparaison après traitement).
- Décision de chaînage par rapport à un seuil de poids.

	Nom	Prénom	DdN	Total
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

Exemple

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		15.2	96.68	3.32
1	1	1	8 852	Seuil chaînage	24.4	99.99	4e-04

$P(m)$: Proba. que les 2 enregistrements de la paire correspondent au même individu.

$G(u)$: Proba. que les 2 enregistrements correspondent à 2 individus différents.

Seuils d'appariement

On chaîne les couples dont les poids composés sont supérieurs à 10 (ad hoc).
La décision finale est **fonction du contexte de l'étude** ; on souhaite :

- un chaînage exhaustif ? $\Rightarrow$ baisser le seuil, donc accepter des faux positifs ou pratiquer une validation
  - automatique ; exemple : les dates de distribution PSL compatibles avec dates d'hospitalisations,
  - ou manuelle ; exemple : retour au dossier médical pour vérifier le diagnostic (cancer, ...).
- identification de vrais positifs ?

Choix des variables

Les variables les plus utiles à garder pour l'appariement :

- Dépend des variables disponibles et de l'étude (ex : date de naissance chez nouveau-nés VS pop. générale),
- Celles ayant la meilleure valeur discriminante (à estimer en fonction de la population étudiée et de la qualité des données),
- Sexe non pertinent (faible valeur discriminante), mais permet de conforter/valider le chainage, pas dans le modèle,
- Commune de résidence, oui si le nombre de communes est important et si les deux sources de données concernent les mêmes dates (pas de déménagement entre temps).

Utilité ou non des traitements phonétiques

- Dépend de la qualité des données.
- Sur nos données, il nous a paru préférable de ne pas appliquer de traitement phonétique car l'algorithme de chainage permet de retrouver les cas douteux sans perte d'information. Appliquer toutefois une normalisation (accents, ponctuation, majuscules/minuscules, etc..).
- Attention, risque de collisions si l'on applique des traitements phonétiques.

Conclusion sur le chaînage

S'il n'y a qu'une seule chose à retenir, c'est que tout (champs identifiants, seuils) dépend du contexte de l'étude :

- besoins,
- données et leur qualité,
- possibilités de vérifications,
- ...

De plus, d'autres traitements peuvent avoir lieu :

- en amont pour limiter la taille du produit cartésien (*blocking*),
- en aval pour conforter les appariements, en utilisant d'autres champs (sexe, commune, ...) non utilisés pour le chaînage avec le modèle.

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).

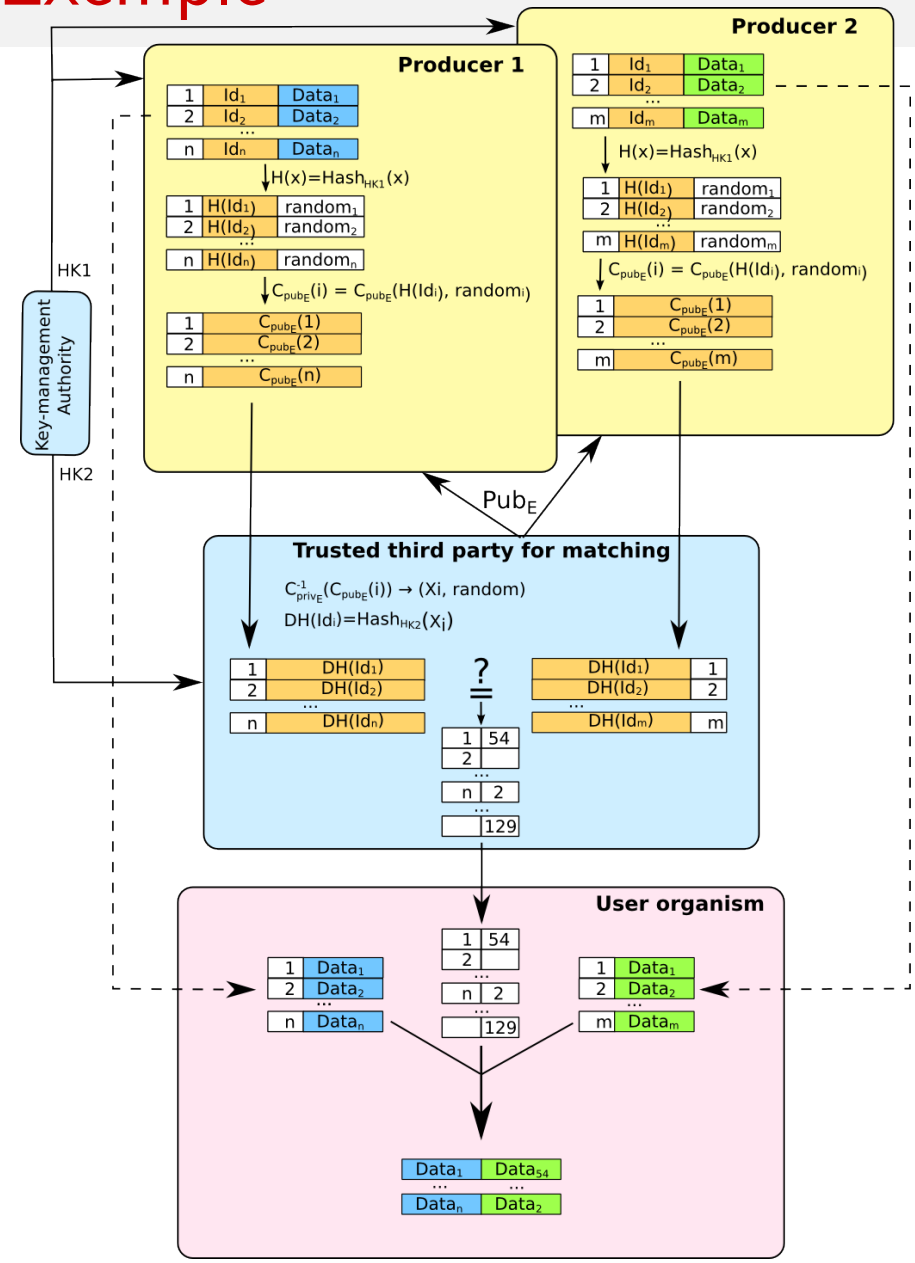
Organisation

Pour protéger l'anonymat (ou un certain niveau d'anonymat, puisqu'on ne sait jamais de quelles données supplémentaires dispose un tiers) lors d'appariements de données :

- compartimenter les données, en séparant les données identifiantes des autres données,
- ceci afin de maîtriser qui peut apparier quelles informations,
- pour toujours s'assurer de l'impossibilité de ré-identifier des individus.

Il ne s'agit plus de technique pure mais d'organisation des circuits d'information.

Organisation - Exemple



Contrôle d'accès

Pour la manipulation de données qui doivent rester précises et donc comportent des risques de réidentification, exemple du CASD :

≪ Le centre d'accès sécurisé aux données (CASD) du Genes (Groupe des écoles nationales d'économie et statistique) est un équipement conçu pour permettre aux chercheurs de travailler sur des données individuelles très détaillées, et donc soumises à la confidentialité, dans des conditions de sécurité élevées.≫

Il s'agit d'offrir un accès aux données à distance via un équipement spécifique (un terminal connecté de façon sécurisée au CASD). La sortie et l'entrée de données font l'objet de procédures et sont réalisées à la demande du chercheur par le CASD.

Le CASD donne accès aux données de différents fichiers et enquêtes publiques.

Loi sur le numérique

- L'occasion de permettre/faciliter des études statistiques, notamment pour la recherche...
- ... tout en protégeant la vie privée.
- Mieux vaut mettre des barrières (compartimentation des données) et ne pas compter que sur des autorisations ou interdictions dans le droit.

"La confiance n'exclut pas le contrôle."

Merci