Gestion des données médicales anonymisées : problèmes et solutions

Anas Abou El Kalam*, Yves Deswarte *, Gilles Trouessin**, Emmanuel Cordonnier ***

* LAAS-CNRS. 7 avenue du colonel Roche – 31077 Toulouse Cedex 4 .{Deswarte, anas}@laas.fr.
** ERNST & YOUNG. 1 place Alfonse Jourdain – 31000 Toulouse, gilles.trouessin@fr.ey.com.
*** ETIAM. 20 Rue du Pr Jean Pecker – 35000 Rennes, emmanuel.cordonnier@etiam.com.

Résumé

L'anonymisation des identités des personnes figurant dans des fichiers informatisés, en particulier lorsqu'ils contiennent des informations sensibles pouvant porter atteint à la vie privée, est une préoccupation actuelle majeure. L'étude des techniques d'anonymisation que nous présentons ici s'articule comme suit :

Après avoir défini la problématique, les procédures d'anonymisation les plus typiques sont présentées et analysées. Un ensemble de concepts relatifs à l'anonymisation ainsi qu'une démarche rigoureuse d'analyse des besoins et de choix des solutions sont ensuite proposés. Cette démarche est essentiellement fondée sur l'identification des besoins, des objectifs et des exigences de sécurité, afin de définir ou de choisir la solution la plus adaptée à chaque problème lié au respect de la vie privée.

L'éventail des cas d'anonymisation pouvant exister dans le domaine médical est ensuite décrit, et un ensemble de scénarios possibles est identifié et confronté à la démarche d'analyse présentée.

Enfin, une nouvelle procédure générique d'anonymisation des données identifiantes et de chaînage des informations est proposée. Une partie de cette procédure est réalisée au sein d'une future possible génération de cartes qui contiendrait un identifiant anonyme (généré aléatoirement) du patient et qui supporterait l'exécution de traitements simples comme MD5 ou SHA. Pour réaliser l'anonymisation, les bases de données subissent des transformations cryptographiques à différents niveaux : en amont au sein des hôpitaux ou des centres de soins, puis en aval dans des centres de traitement.

La procédure présentée dans cet article répond aux exigences des législations françaises et européennes qui protègent les droits et la vie privée des individus : elle tient compte de la finalité des traitements, garantit la prise en compte du consentement du patient, résiste aux attaques par dictionnaire, respecte le principe du moindre privilège, etc. Pour autant elle reste flexible, adaptable à différents secteurs, et supporte des changements organisationnels tels que la fusion de plusieurs hôpitaux.

Mots clés

Systèmes d'informations médicales, sécurité des systèmes d'information, protection de la vie privée, anonymisation, démarche d'analyse et d'expression des besoins de sécurité, PMSI.

1. Problématique

Bien que l'instauration du réseau de soins facilite la communication des données entre différentes structures, elle pose des problèmes concrets de sécurité. D'une part, l'usage d'informations médicales nominatives pour les soins impose d'avoir l'assurance de l'identité des personnes auxquelles se rapportent ces informations. D'autre part, en facilitant l'échange, le partage et le traitement des données de santé entre plusieurs acteurs, la réidentification du patient à partir d'informations anonymisées accessibles peut contribuer à briser le secret médical ou à permettre d'inférer des informations confidentielles, et donc porter atteinte à la vie privée des patients.

Malheureusement, il est souvent possible d'identifier un individu par un simple rapprochement de données personnelles de nature médicale ou sociale. Par exemple, l'âge, le sexe et le mois de sortie de l'hôpital, permettent d'isoler le patient dans une population restreinte ; la donnée de deux dates (voire de deux semaines) d'accouchement pour une femme permet de l'isoler dans une population plus grande (typiquement, la population d'un pays comme la France).

Pour faire face à des problèmes de ce type, les législations internationales [Résolution 1990] et européenne [Directive 2002 ; Directive 1995 ; Recommandation 1997] visent à protéger les *données personnelles* et interdisent tout croisement de fichiers. De même, en France, la loi « informatique et libertés » [Loi 1978] accorde une protection particulière aux *données nominatives*. Il existe toutefois une certaine nuance entre donnée nominative et donnée à caractère personnel : même après anonymisation, une donnée nominative peut rester à caractère personnel, et il est parfois possible, de réidentifier des données anonymisées.

La base législative que nous venons de citer doit être traduite, au niveau du système informatique, par des mécanismes de sécurité. A cet égard, et selon les besoins, le choix en terme de protection peut faire appel à des solutions techniques comme le hachage et le chiffrement, ou à des solutions organisationnelles comme les politiques de contrôle d'accès ou les anonymisations thématiques.

À cet égard, un premier niveau de confidentialité peut être assuré en chiffrant les données transmises, de façon à ce qu'elles ne soient déchiffrées que par le ou les destinataires légitimes. Ce peut être le cas d'un échange de données médicales entre un laboratoire d'analyses médicales et un médecin traitant.

Dans d'autres cas, on souhaite garder l'anonymat de certaines données (en particulier les noms, prénoms, adresses, ou le numéro de sécurité sociale) même si le destinataire est légitime. Par exemple, il faut garder l'anonymat des patients lorsque leur cas est présenté dans des publications scientifiques, ou lors de la transmission des informations relatives à l'activité des médecins libéraux vers les unions professionnelles[1].

À l'inverse, si le but est de pouvoir analyser des trajectoires de soins, c'est-à-dire, avoir un suivi permanent des évolutions des maladies, on doit respecter l'anonymat (des patients) tout en ayant la possibilité de chaîner les données qui concernent un même patient. Il convient ainsi d'utiliser un identifiant anonyme, mais toujours le même pour un patient donné.

Enfin, il est parfois souhaitable qu'une autorité puisse croiser les données anonymisées avec d'autres données anonymes concernant le même individu, ou même lever l'anonymat dans des cas bien particuliers. Par exemple, dans le cas des études épidémiologiques, les corrélations entre plusieurs pathologies peuvent nécessiter de remonter aux identités réelles pour compléter a posteriori les données recueillies antérieurement, et ainsi affiner ces études.

2. L'anonymisation dans les pays européens

2.1 L'anonymisation dans la région Bourgogne

En 1995, le CHU de Dijon, ainsi que d'autres établissements de santé de la région Bourgogne ont choisi l'algorithme de hachage SHA (*Standard Hash Algorithm*) pour transformer, d'une manière irréversible, les informations d'identification : nom, prénom, date de naissance et sexe. Le but est d'obtenir un identifiant strictement anonyme, mais toujours le même pour un patient donné [Quantin *et al*. 1998]. Pour illustrer cette procédure d'anonymisation, considérons le cas où les hôpitaux transmettent des données médicales au Département d'Informations Médicales (DIM).

Si on n'utilise qu'une fonction de hachage du côté des hôpitaux, les personnes en charge des traitements statistiques et médico-économiques (au DIM) peuvent remonter aux identités par une simple attaque ponctuelle ou par dictionnaire. En effet, ils peuvent appliquer l'algorithme de hachage (supposé public) à un certain nom (Paul Dupont, par exemple) et obtenir le code anonyme (123, par exemple) associé à ce nom ;

[1] Une union professionnelle peut être définie comme une assemblée de médecins élus par région ou regroupés en sections.

puis le comparer avec les données (123 – SIDA …, par exemple) reçues des hôpitaux, pour enfin établir le lien entre les informations médicales (SIDA) et le nom (Paul Dupont).

De la même manière, si on n'utilise qu'une fonction de hachage du côté du DIM (avant l'archivage), les employés des hôpitaux pourront retrouver les données (utilisés dans les archives) de n'importe lequel de leurs patients, y compris celles provenant d'autres établissements.

Pour empêcher ce type d'attaques, deux clés ont été ajoutées à l'algorithme de hachage SHA. L'identité est d'abord concaténée à une première clé k_1, utilisée par tous les émetteurs des données (hôpitaux et médecins). Une fonction de hachage est ensuite appliquée au résultat : $empreinte_1 = H(k_1 \mid identité)$. Cette opération produit une empreinte qui varie d'une identité à l'autre, mais qui est toujours la même pour un patient donné. Les informations transmises au centre de traitement des fichiers (DIM) en vue de leur rapprochement sont ainsi devenues strictement anonymes et les personnes qui assurent les traitements centralisés ne peuvent pas lever l'anonymat à l'aide d'une attaque par dictionnaire puisqu'elles ne connaissent pas la clé k_1. De l'autre côté de la communication, les informations reçues par le DIM sont hachées par le même algorithme mais avec une seconde clé k_2, qui n'est pas communiquée aux hôpitaux : $empreinte_2 = H(k_2 + empreinte_1)$ (voir figure 1).

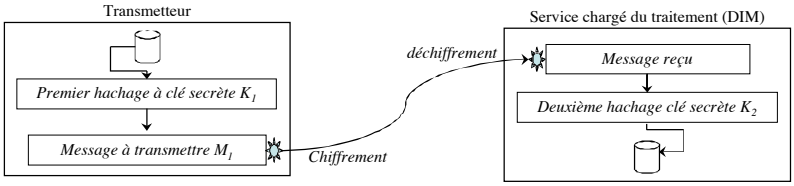

Figure 1 : Procédure de double hachage des informations traitées par le DIM.

Il est clair que ce protocole s'avère complexe et risqué. En effet, il nécessite une distribution de la même clé secrète k_1 à tous les fournisseurs d'informations (médecins libéraux, hôpitaux, cliniques, etc.), alors que la confidentialité de cette clé est primordiale. Si une clé (k_1 ou k_2) est corrompue, le niveau de sécurité est considérablement réduit. De même, si un jour il s'avère que l'algorithme (ou la longueur de la clé) n'est plus efficace, comment faire le rapprochement entre les identifiants avant et après changement de l'algorithme ou de la clé ? Si ce problème survient, la seule solution envisageable consiste à appliquer une double transformation à toute la base de données, solution qui n'est guère aisée.

2.2 La procédure FOIN

La procédure FOIN (Fonction d'Occultation d'Informations Nominatives) a été élaborée par le CESSI (Centre d'Etudes des Sécurités des Systèmes d'Information) de la CNAM-TS (Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés) pour le PMSI (Programme de Médicalisation des Systèmes d'Information) privé. Comme en Bourgogne, FOIN utilise une fonction de hachage à sens unique (SHA) avec une clé des deux côtés. L'originalité de cette méthode réside dans l'utilisation de la technique de *Fragmentation-Redondance-Dissémination* ou FRD [Fabre *et al*. 1996]. Les étapes de cette technique de tolérance aux intrusions sont les suivantes :

- découper l'information (clé secrète) en fragments de telle sorte que des fragments isolés ne puissent fournir d'information significative ;
- ajouter de la redondance pour empêcher que la modification ou la destruction de quelques fragments n'ait pas de conséquence pour les utilisateurs autorisés ;
- isoler les fragments les uns des autres par dissémination de sorte qu'une intrusion (la corruption d'une partie de la clé secrète) dans une partie du système ne fournisse que des fragments isolés.

Ainsi, FOIN fragmente la clé secrète en *N* images, à l'aide d'un schéma à seuil de Shamir [Shamir 1979], de telle sorte que la clé ne peut être reconstruite qu'à partir d'un certain nombre *s* (dit seuil de reconstitution) d'images différentes. Les images de la clé secrète sont disséminées sur un nombre de supports distincts. La première image sera placée dans la fonction d'anonymisation (dans le logiciel distribué aux

transmetteurs d'informations), les autres sont données à des personnes de confiance, comme le responsable de l'application ou le directeur de la CNAM. Ainsi, même s'il existe *N* images (fragments) de la clé, la présence de "*s*-1" personnes de confiance est suffisante pour reconstituer le secret (figure 2).

Comme en Bourgogne, et pour se prémunir contre toute attaque ponctuelle ou par dictionnaire, la fonction FOIN est utilisée à deux niveaux : une première fois dans les hôpitaux, avant de transmettre les données médicales des patients, et une deuxième fois, avant l'archivage de ces données.

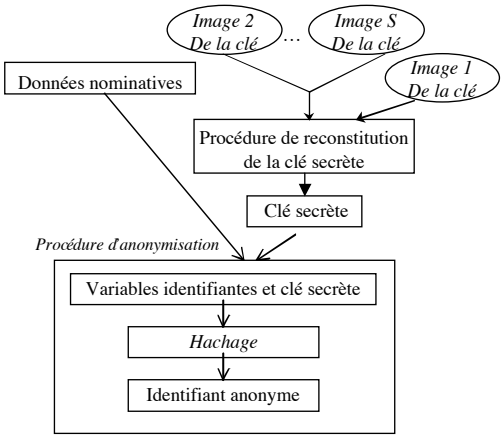

Figure 2 : Procédure FOIN.

Néanmoins, le cas où *s* = 2 est confronté aux mêmes faiblesses que la procédure du CHU de Dijon. Si en plus, *N* > 2, les images seront détenues par *N* personnes et n'importe laquelle pourra reconstituer la clé (puisqu'une des deux images nécessaires est présente dans le logiciel). De plus, la technique de FRD ne résout que le problème du stockage "longue durée" ; la clef reste vulnérable au vol par un utilisateur malveillant qui réussit à la lire ou la copier lorsqu'elle est utilisée durant les traitements.

2.3 Le traitement statistique des données médicales en Suisse

Du point de vue statistique, il n'est pas nécessaire de savoir à qui appartient un fichier médical. Néanmoins, en Suisse, l'Office Fédéral des Statistiques (OFS) a besoin de reconnaître si deux fichiers différents correspondent à la même personne. L'implémentation suisse propose de hacher les identifiants dans les hôpitaux avant de les transmettre à l'OFS ; puis, à la réception, les données médicales sont chiffrées par la clé secrète de l'OFS [Jeanneret *et al*. 2001] (figure 3). Les étapes de cette implémentation peuvent être résumées comme suit :

- Calcul d'une empreinte à partir d'un hachage des informations identifiantes (date de naissance, sexe, nom et prénom) : *Hachage*[*Var-ID*] = *Empreinte*.
- Génération en arrière-plan (dans l'ordinateur de l'hôpital) d'une clé *c* de session.
- Chiffrement de l'empreinte avec IDEA en employant *c* : $IDEA[Empreinte]_c$; et chiffrement de *c* par la clé publique *E* de l'OFS en utilisant l'algorithme RSA : $RSA[c]_E$.
- Transmission de la clé de session (chiffrée), de l'empreinte (chiffrée) et des données médicales (chiffrées) à l'OFS.
- À la réception, déchiffrement de la clé secrète *c* par RSA en employant la clé privée *D* de l'OFS ;

- Déchiffrement de l'empreinte (avec IDEA et la clé *c*), et re-chiffrement de cette empreinte avec une clé *k* (fragmentée) pour donner le lien anonyme utilisé comme code personnel (code de liaison uniforme) lors du stockage des données médicales au niveau de l'OFS.

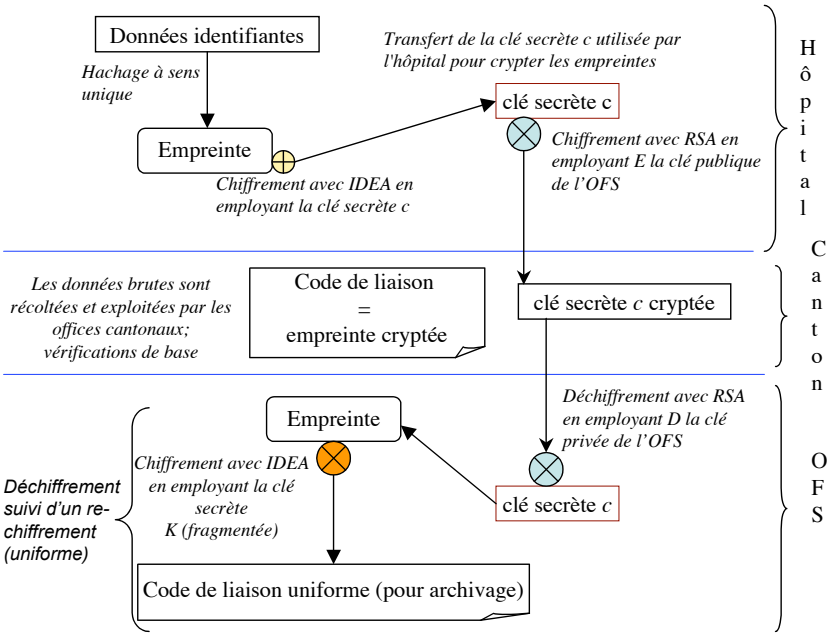

Figure 3 : Transformation des données identifiantes en Suisse.

Les opérations effectuées au niveau de l'OFS ne doivent jamais être visibles aux utilisateurs de l'OFS. Cependant, comment s'assurer qu'elles ne sont jamais enregistrées sur aucun support ? Un cheval de Troie opérant pour une personne malveillante pourrait récupérer les valeurs de la clé secrète *c* ou des *empreintes*, et effectuer par la suite une attaque par dictionnaire. Pour pallier ce type de risques, il faut que ces étapes (phase de *calcul*) soient exécutées par un module matériel bien protégé. Des mécanismes inviolables de contrôle d'accès, éventuellement matériels, pourront renforcer la protection de ce module de façon à ce que seules les personnes de confiance puissent réaliser l'opération composite *calcul* ; le serveur d'autorisation leur donne les droits correspondants sans qu'elles puissent lire ou copier, ni l'*empreinte*, ni les clés secrètes *K* et *c* ; dans ce cas, des droits distribués sont donnés aux différents composants matériels ; chaque composant effectue les opérations qui lui sont destinées.

3. Démarche d'analyse

L'analyse menée dans la section précédente montre, les avantages et les faiblesses de chacune des solutions actuelles et renforce notre volonté d'une démarche analytique préalable des risques, des besoins, des exigences ainsi que des objectifs de sécurité, avant de recourir aux solutions de sécurité assurant l'anonymisation. Mais avant tout, définissons quelques notions essentielles.

Le terme anonymisation recouvre deux grandes catégories de concepts :

- la demande sous forme de besoins d'anonymisation à satisfaire ;
- la réponse sous forme de fonctionnalités et solutions pour anonymiser.

Comme beaucoup de fonctionnalités liées à la sécurité, une fonctionnalité d'anonymisation peut être exprimée selon un des trois niveaux d'attente suivants :

- le *besoin* d'anonymisation, qui représente les attentes de l'utilisateur ; généralement sous une forme qui n'est pas toujours très explicite ni aisée à formaliser ;
- l'*objectif* d'anonymisation, qui spécifie le niveau de sécurité à atteindre ou les menaces à éviter (comment satisfaire les exigences ?).

- l'*exigence* d'anonymisation, qui représente la façon d'exprimer le besoin ; dans la mesure du possible, très proche d'un formalisme clair et d'une sémantique non-ambiguë.

Sachant que les besoins ont déjà été identifiés dans la problématique (première section), commençons par identifier ce que nous entendons par objectifs et exigences d'anonymisation. La fin de la section suivante revient en détail sur l'ensemble des besoins à travers un ensemble de scénarios, et identifie, pour chaque scénario, ses objectifs et ses exigences d'anonymisation.

3.1 Objectifs d'anonymisation

Un objectif d'anonymisation est défini en fonction de l'une des trois propriétés suivantes applicables à la fonction d'anonymisation [Trouessin 2001] :

- *réversibilité* : cacher les données par un simple chiffrement des données. Dans ce cas, il y a possibilité de remonter depuis les données chiffrées jusqu'aux données nominatives originelles.
- *irréversibilité* : c'est le cas réel de l'anonymisation ; une fois remplacés par des identifiants anonymes, les identifiants nominatifs originels ne sont plus recouvrables ; cependant, avec les techniques d'attaques par inférence[2], les identifiants anonymes, s'ils sont trop universellement utilisés, risquent de permettre la découverte d'identités mal cachées ; pour ce type d'anonymisation, la technique communément utilisée est une fonction de hachage ;
- *inversibilité* : c'est un cas où il est impossible en pratique de remonter aux données nominatives, sauf en appliquant une procédure exceptionnelle, sous le contrôle d'une instance légitime (médecin-conseil, médecin inspecteur), garante du respect de la vie privée des individus concernés ; il s'agit cette fois-ci d'une pseudonymisation au sens des critères communs[3] [CC 1999].

3.2 Exigences d'anonymisation

Des informations sur l'environnement informatique (utilisateurs, attaques, etc.) du système étudié permettent de compléter l'analyse du besoin. En l'occurrence, même si les informations sont anonymes, un utilisateur malveillant peut construire divers types de raisonnement pour déduire des informations confidentielles. Les exigences d'anonymisation sont exprimées en terme de *chaînage* (continuité de l'anonymisation) et de *robustesse* (sûreté de l'anonymisation).

Le *chaînage* permet d'associer un ou plusieurs identifiants anonymes à une même personne physique. Comme indiqué sur la figure 4, un chaînage peut être temporel (toujours, parfois, jamais) ; géographique (international, national, régional, local) ; ou spatio-temporel (par exemple, "toujours et partout", "parfois et partout", "local et jamais") [AFNOR 1997].

La *robustesse* d'un système d'anonymisation est constituée de l'ensemble des caractéristiques à satisfaire vis-à-vis d'attaques ayant pour but de lever l'anonymat de façon non-autorisée. Il peut s'agir d'une *robustesse à la réversion* concernant la possibilité d'inverser la fonction d'anonymisation, mais il peut aussi s'agir d'une robustesse à l'inférence qui consiste à déterminer des informations nominatives à partir d'éléments d'informations purement anonymes. En général, une inférence peut être :

- *déductive* : elle utilise la logique du premier ordre (valeurs : oui, non ; opérateurs : et, ou) pour déduire des informations confidentielles non accessibles ; par exemple, si un certain patient fait un test de dépistage *puis* dans les quelques jours qui suivent, fait un test de dosage, *alors* le résultat du dépistage était positif ;
- *inductive* : elle s'apparente alors à des raisonnements de type loi des grands nombres sans forcément l'appliquer sur une grande échelle ; cela consiste, par exemple, à induire qu'un patient est très certainement atteint de telle pathologie compte tenu du fait qu'il lui est prescrit tels médicaments comme il est d'usage pour cette pathologie ;

[2] Une inférence est la découverte de données confidentielles non directement accessibles, rendue possible par la mise en correspondance de plusieurs données légitimement accessibles, révélant tout ou partie des informations relatives à une personne.

[3] Selon les critères communs, une pseudonymisation est une anonymisation où la personne concernée peut être tenue pour responsable de ses actes.

- *abductive* : lorsqu'un raisonnement classique utilisant les informations explicitement stockées dans le système d'informations ne permet pas d'inférer d'informations, mais ce raisonnement pourrait être complété en faisant des hypothèses sur certaines informations, par exemple, "et s'il avait un cancer, cela expliquerait pourquoi il s'absente du Conseil des Ministres pour se rendre à l'hôpital Paul Brousse de Villejuif …".
- *probabiliste* (*ou adductive*) : elle parvient à estimer la vraisemblance d'une information sensible en utilisant les informations accessibles, par exemple, "puisque *P* est traité à l'hôpital *H*, et puisque *H* est spécialisé dans les maladies M_1 et M_2, et puisque, à son âge, la probabilité d'avoir M_1 est très faible (10%), alors on peut déduire qu'à 90%, *P* est atteint de M_2".

Cette liste n'est pas exhaustive et on peut naturellement imaginer d'autres types de canaux d'inférence fondés sur d'autres types de raisonnement, tel que le raisonnement par évidence ou par analogie.

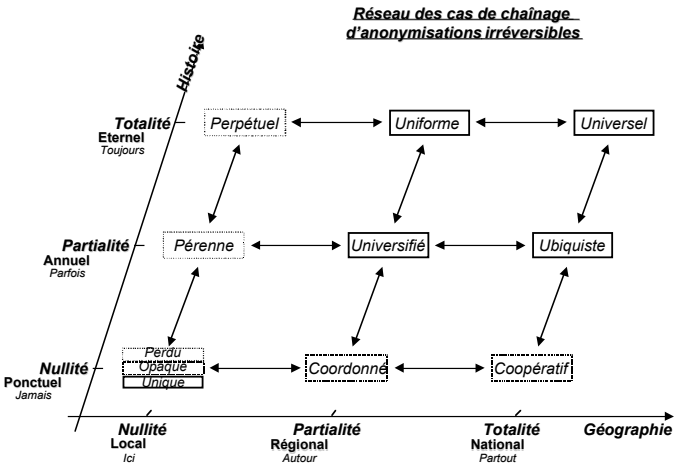

Figure 4 : Anonymisation en cascade : de l'universalité jusqu'à l'unicité.

3.3 Analyse de scénarios du domaine santé-social

Lors du transfert des données médicales

La sensibilité des informations échangées entre professionnels de santé (par exemple, le laboratoire d'analyses et le médecin) met en évidence le besoin de confidentialité et d'intégrité des données transitant sur le réseau de soins. La figure 5 schématise une des solutions qui consiste à utiliser un chiffrement asymétrique. Ainsi, en supposant que le destinataire légitime est le seul à disposer de la clé privée, personne d'autre ne peut déchiffrer le message transitant par le réseau et ainsi accéder aux données personnelles en clair. Si les données sont volumineuses, il est préférable d'utiliser un chiffrement hybride : une clé symétrique de session est générée aléatoirement, et utilisée pour chiffrer (par un chiffre symétrique) le message, puis cette clé symétrique est chiffrée avec la clé publique du destinataire ; le message chiffré et la clé de session chiffrée sont alors envoyés au destinataire.

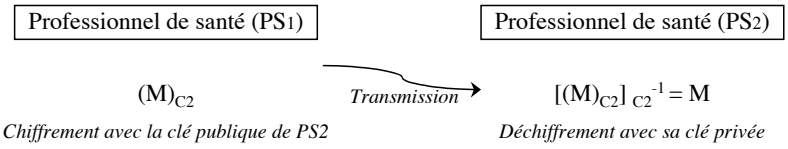

Figure 5 : Échange de données chiffrées entre professionnels de santé.

Selon la classification donnée précédemment, l'*objectif* est une anonymisation réversible, tandis que l'*exigence* est la robustesse à l'inversion.

Notons que le chiffrement des données médicales transitant sur le réseau est actuellement une pratique de plus en plus répandue entre les professionnels de santé, notamment en utilisant S-MIME.

Pour les unions professionnelles

Le transfert des données relatives aux activités des médecins vers les unions professionnelles se fait à des fins d'évaluation de l'activité des médecins. Une première exigence consiste donc à cacher les identités du patient et du médecin. Toutefois, l'anonymat des médecins doit pouvoir être levé pour l'évaluation de leurs comportements en vue de la qualité de soins. L'analyse de l'article L4134-4 du code de la santé publique ainsi que l'article 81 de la loi 94-43 [Loi 1994] nous permet de déduire les objectifs suivants de sécurité :

- l'anonymisation inversible de l'identité du médecin ; seule une autorité habilitée à évaluer les comportements des médecins pourrait rétablir les identités réelles des médecins ;
- l'anonymisation inversible des données nominatives du patient, seuls les médecins-conseils de la sécurité sociale pourront lever cet anonymat ; en effet, l'article L161-29 du code de la sécurité sociale ajoute : « seuls les praticiens-conseils et les personnels sous leur autorité ont accès aux données nominatives (des patients) issues du traitement susvisé, lorsqu'elles sont associées au numéro de code d'une pathologie diagnostiquée ».

Cette manière de faire évite les risques suivants (au niveau des unions professionnelles) :

- un utilisateur malhonnête qui tente d'avoir plus de détails sur les activités d'un médecin alors que la finalité de son traitement ne le justifie pas ; par exemple, dans le cadre d'une étude relative au fonctionnement du système de santé, il n'est pas nécessaire d'accéder aux identités (respect du principe du moindre privilège) ;
- atteinte à l'intimité des patients dans la mesure où ceux-ci peuvent confier des informations à certains professionnels de santé, sans pour autant avoir forcément envie de les communiquer aux autres professionnels de santé ou personnes en charge des traitements au sein des unions.

Le scénario décrit dans cette section est résumé dans la figure 6.

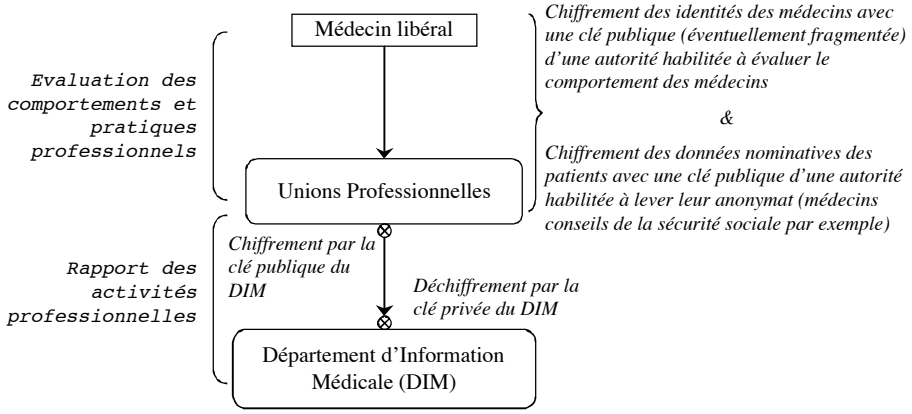

Figure 6 : Manipulations des identités au niveau des unions professionnelles.

Dans le cadre du PMSI

Le Programme de Médicalisation des Systèmes d'Information (PMSI) est un système d'analyse de l'activité des établissements de santé dont la finalité est l'allocation des ressources tout en diminuant les inégalités budgétaires. Le PMSI a été expérimenté depuis 1983, et généralisé dans les hôpitaux publics et privés participant au service public par la circulaire du 24 juillet 1989 [Circulaire 1989] pour l'activité de MCO (Médecine, Chirurgie, Obstétrique). Son utilisation à des fins budgétaires a été formalisée par la circulaire du 7 décembre 1996 [Ordonnance 1996]. Il a été étendu aux établissements privés par les ordonnances du 24 avril 1996. La circulaire du 9 mars 1998 [Circulaire 1998] a généralisé le PMSI aux établissements publics ayant une activité de soins, de suite et de réadaptation. Une multitude de textes ont été élaborés pour réglementer le fonctionnement du PMSI. Citons à titre indicatif, la loi du 31 juillet 1991 [Loi 1991]le décret du 27/07/94 ainsi que les arrêtés des 20/09/1994, 22/07/1996 et 29/07/1998.

Dans la pratique, chaque séjour d'un patient donne lieu à un recueil standardisé de données de nature administrative (dates d'entrée et de sortie, date de naissance, nom et prénom) et de nature médicale (diagnostics, actes codés). Les séjours sont ensuite classés selon l'indicateur médico-économique "Groupe Homogène de Malades" (GHM). Les patients d'un GHM donné sont considérés comme ayant mobilisé des ressources de même ampleur. Chaque année une échelle des coûts affecte un coût relatif à chaque GHM, mesuré en points ISA pour "Indice Synthétique d'Activité". Les données du PMSI des établissements publics sont anonymisées, puis transmises semestriellement aux Agences Régionales de l'Hospitalisation (ARH) qui les utilisent pour l'allocation budgétaire. Celles des établissements privés sont transmises trimestriellement à la CNAM-TS, en attendant de devenir un outil d'allocation de ressources. Plus précisément, tout séjour hospitalier effectué dans la partie court séjour d'un établissement fait l'objet d'un Résumé de Sortie Standardisé (RSS), constitué d'un ou plusieurs Résumés d'Unité Médicale (RUM). Le RUM contient des données (administratives et médicales) concernant le séjour d'un patient dans une unité médicale. À partir des RUM, le Département d'Information Médicale (DIM) construit le fichier des Résumés de Sortie Standardisés (RSS) à l'aide d'un logiciel regroupeur. Les services des statistiques et des études épidémiologiques reçoivent du médecin du DIM, les données médicales et administratives figurant sur les Résumés de Sortie Anonymisés (RSA). La procédure générale est donnée sur la figure 7.

Étant donné que la finalité du PMSI est purement médico-économique (et non pas directement épidémiologique), le *besoin* est de pouvoir effectuer des trajectoires de soins par le biais d'une pseudonymisation ; l'*objectif* est une anonymisation irréversible ; et les *exigences* sont un chaînage universel (toujours et partout le même identifiant pour un patient donné) ainsi que la robustesse à la réversion et aux inférences (déductives, inductives, abductives, etc.).

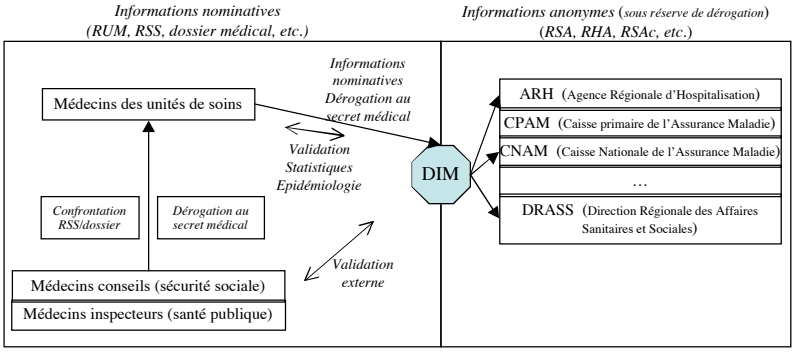

Figure 7 : Frontières des données nominatives, anonymes et anonymisables.

Gestion des données sur les maladies à déclaration obligatoire

Les maladies dont la surveillance est nécessaire à la conduite et à l'évaluation de la politique de santé publique (le SIDA, par exemple), ou qui nécessitent une intervention urgente locale (méningite, choléra, rage) sont des maladies à déclaration obligatoire. À l'origine, les fichiers des patients séropositifs sont nominatifs, mais ils sont anonymisés (*anonymisation irréversible*) avant toute transmission.

Les *besoins* sont divers : prévention, production de soins, veille sanitaire, analyses épidémiologiques, etc. L'objectif principal est l'irréversibilité de la fonction d'anonymisation. Le chaînage universel et la robustesse à la réversion et aux attaques par inférence constituent les principales exigences.

À cet égard, le type de protection doit dépendre des objectifs. En effet, s'agit-il d'obtenir, année par année, un état exhaustif du nombre de séropositifs pour connaître l'évolution de l'épidémie, ou d'évaluer, de façon globale, l'impact des actions de prévention ? S'agit-il encore d'instituer une véritable surveillance épidémiologique de l'évolution des cas d'infection par le VIH, du stade de la découverte de la séropositivité, à l'apparition éventuelle du SIDA avéré ? Dans ce cas, l'objectif est de mesurer de façon fine l'impact des actions thérapeutiques et de prévention nécessitant un suivi des cas.

Ce choix d'objectifs comporte des conséquences importantes tant sur la nature des données susceptibles d'être collectées que sur leur durée de conservation et les liens éventuels avec d'autres systèmes de surveillance. Il implique en conséquence des choix en terme de protection de données.

Appliquer l'anonymisation à la source et disposer de mesures de sécurité adéquates ne dispense pas de s'interroger sur la pertinence des autres informations figurant sur la déclaration de séropositivité. Il s'agit en particulier, du code postal de résidence, la profession et l'origine géographique.

- Le code postal de domicile : si l'objectif est de mieux cibler les actions de prévention locale, sa collecte semble nécessaire. Néanmoins, la pertinence du recueil de cette donnée n'est pas à ce jour réellement démontrée. En outre, sa collecte et son expiration pourraient être de nature à permettre une localisation géographique précise surtout dans les petites communes. Dès lors le recueil sous une forme aussi détaillée que le code postal du lieu de résidence des personnes séropositives peut paraître excessif au regard des objectifs recherchés et il est probablement plus judicieux d'utiliser le code du département au lieu du code postal.
- La profession : il ne paraît pas nécessaire de disposer de la profession précise ; une simple mention des catégories socio-professionnelles paraît être pertinente.
- L'origine géographique : il serait peut-être suffisant de mentionner si la personne est originaire d'un pays où la transmission hétérosexuelle est prédominante ou si elle a eu des relations sexuelles avec une personne ayant vécu dans un pays où la contamination hétérosexuelle est prédominante.

Certes, l'appauvrissement peut contribuer au respect de la vie privé, néanmoins, s'il est trop important, il peut fausser les statistiques et remettre en cause la fiabilité scientifique de la surveillance épidémiologique.

Traitements des données statistiques

En aucun cas, les données médicales à caractère personnel ne peuvent être manipulées pour des traitements à des fins non-épidémiologiques, par exemple, des traitements purement statistiques ou à des fins de publications scientifiques. À cet égard, non seulement ces données doivent être anonymisées, mais il doit être impossible de les ré-identifier. Ainsi, s'imposent l'irréversibilité de l'anonymisation ainsi que la robustesse aux inférences. En effet, même après anonymisation, les identités peuvent être déduites par un statisticien en combinant plusieurs requêtes ou en complétant son raisonnement par des hypothèses ou par des informations externes au système.

Le domaine de l'inférence d'information dans les bases de données a été étudié depuis de nombreuses années, et il a fait l'objet d'une littérature abondante [Denning 1979 ; Cuppens 2003, Castano et al. 1995]. La sécurité dans les bases de données statistiques est un problème réel, et plusieurs solutions apparaissent dans la littérature, mais il est difficile de décider si l'une d'entre elles est vraiment satisfaisante. Par exemple, une solution serait de permuter les valeurs des attributs des *n-uplets* (lignes) de chaque table de la base de sorte que la précision globale de la statistique est conservée, alors que les réponses précises (concernant des personnes identifiées) seront fausses. La difficulté inhérente à cette approche réside dans la recherche des ensembles d'entrées dont les valeurs peuvent être permutées de cette façon. Une autre solution pourrait être le brouillage, qui consiste à modifier les réponses aux requêtes statistiques en y ajoutant du "bruit" aléatoire pour rendre plus difficile le recoupement entre requêtes.

Études épidémiologiques focalisées

Le PMSI traite des informations médico-administratives, économiques et statistiques, afin de réaliser des analyses pertinentes des bases de données régionales et nationales. Les données traitées sont anonymes, et même si elles sont souvent chaînables, il n'y a généralement aucun moyen de lever l'anonymat. À l'inverse, dans d'autres types d'études, il est souvent souhaitable de revenir à l'identité réelle des patients afin d'améliorer la qualité des soins. Prenons à titre d'exemple, certaines études épidémiologiques focalisées : protocoles de recherche en cancer, maladies génétiques rares, etc .

Supposons, par exemple, qu'une de ces études mette en évidence la situation suivante : les patients de la catégorie "*C*" ayant subi certains traitements "T_{avant}" ont une espérance de vie considérablement réduite s'ils ne suivent pas le traitement "$T_{recouvrement}$". Dans de telles situations, il faudra remonter aux identités réelles

pour que les patients puissent profiter de ces résultats. Il s'agit ainsi d'une anonymisation inversible : seules des personnes habilitées peuvent lever l'anonymat (*médecins conseils, médecins inspecteurs, médecins traitants*) et seulement quand c'est nécessaire.

Dans le cas des protocoles de recherche sur le cancer, le processus commence par un typage (stade de la maladie), puis par une identification du protocole correspondant au patient (s'il existe), enfin, selon le protocole, le patient est enregistré dans un registre régional, national, voire international. Les études épidémiologiques et statistiques faites sur ces registres peuvent dégager de nouveaux résultats concernant les patients d'un certain protocole. Dans le but de raffiner les études et faire avancer la recherche scientifique, il est parfois utile de remonter aux identités réelles des patients pour les identifier, faire des recoupements entre plusieurs données déjà recueillies, et les compléter *a posteriori*.

4. Une nouvelle solution générique

4.1 Schéma général

La section précédente préconise que toute anonymisation nécessite une étude préalable judicieuse, identifiant de manière claire et explicite les besoins, les objectifs ainsi que les exigences. Par ailleurs, l'application de cette démarche à un certain nombre de scénarios identifiés nous a permis de proposer une nouvelle solution générique qui satisfait les exigences soulevées.

Afin de décider quelle vue (forme spécifique de données) est accessible par quel utilisateur, notre solution prend en considération le rôle que joue cet utilisateur, son établissement de rattachement ainsi que la finalité du traitement que subiront les données de cette vue. Bien entendu, ceci respecte le principe du moindre privilège[4] et met en œuvre les recommandations de la norme européenne [CEN 1999].

Pour cela, et comme indiqué sur la figure 9 et détaillé dans la suite de cette section, plusieurs traitements et transformations cryptographiques sont effectués au niveau des hôpitaux, en amont et en aval des centres de traitements (avant la distribution aux utilisateurs finaux).

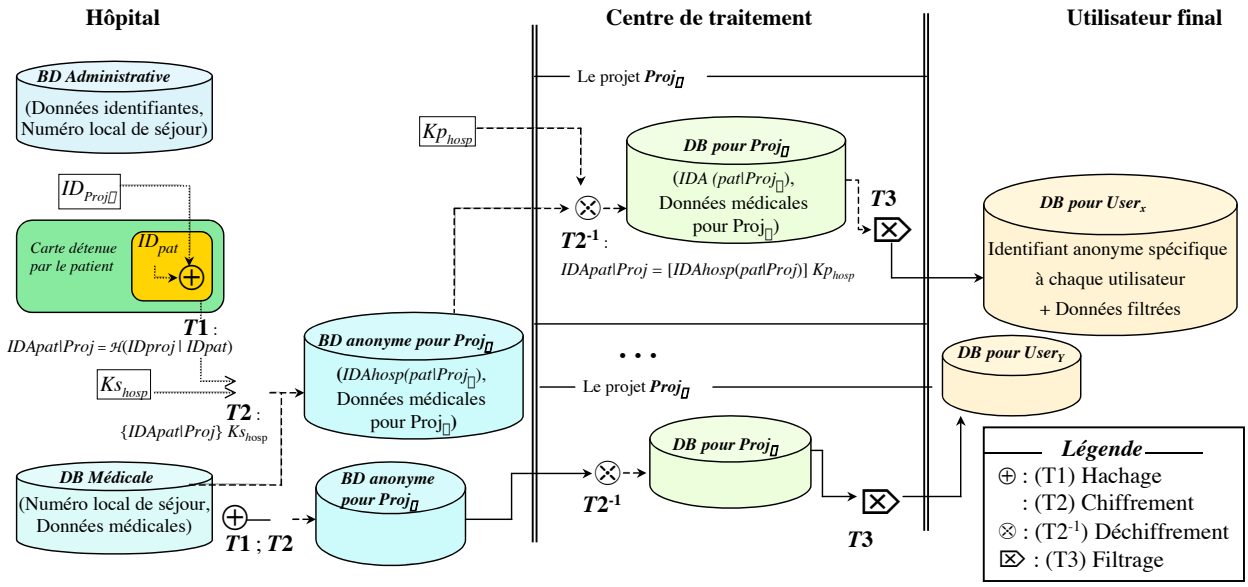

Figure 9 : Procédure d'anonymisation proposée.

[4] Le principe du moindre privilège impose que tout utilisateur ne doit pouvoir accéder à un instant donné qu'aux informations et services strictement nécessaires pour l'accomplissement du travail qui lui a été confié.

Transformations au niveau des établissements de soins

À l'hôpital, trois types de bases de données peuvent être distinguées :

- une base de donnée administrative accessible par les personnels administratifs, chacun selon ses fonctions,
- une base de donnée médicale dont l'accès est restreint aux personnels soignants en charge des patients, ainsi que
- des bases de données anonymes, dont chacune contient les informations nécessaires et suffisantes pour un projet donné. Un projet désigne une entité de traitement des données anonymes tel le PMSI, le DIAM (Dispositif Informationnel de l'Assurance Maladie), les associations de personnes diabétiques, les centres des études cliniques, etc.

Le passage de la base de données médicale à une base anonyme (destinée à un certain projet) nécessite l'application de deux transformations, T1 et T2, aux données à transférer.

La transformation **T1** : consiste à obtenir "$IDA_{pat|Proj}$", un identifiant anonyme par personne et par projet, à partir des deux identifiants :

- "ID_{proj}", l'identifiant du projet, qui est détenu par les établissements de soins (hôpitaux, cliniques) ;
- "ID_{pat}", l'identifiant anonyme unique et individuel du patient ; nous suggérons que cet identifiant soit détenu par le patient dans une carte à puce (type carte VITALE[5]) ; ID_{pat} est généré aléatoirement et n'a aucun lien avec le numéro de sécurité sociale ; une longueur de 128 bits nous paraît suffisante pour éviter des collisions (risque que deux personnes différentes aient le même identifiant).

Au niveau de l'hôpital, et lors de l'alimentation des bases de données anonymes (par projet), l'utilisateur (employé de l'hôpital par exemple) envoie ID_{proj} (l'identifiant du projet concerné par la base de donnée) à la carte du patient ; celle-ci contient déjà ID_{pat} (l'identité du patient donnant son consentement pour l'exploitation de ses données médicales par le projet). La procédure T1 consiste à appliquer une fonction de hachage (MD5 ou SHA par exemple) à (ID_{proj} | ID_{pat}), la concaténation de ID_{proj} et ID_{pat} :

$$\textbf{(T1)} \qquad IDApat|Proj = H(IDproj \mid IDpat)$$

La transformation T1, réalisée au sein de la carte du patient, et produisant l'empreinte $H(ID_{proj} \mid ID_{pat})$, vise les objectifs suivants :

- un patient n'apparaît dans une base de donnée anonyme que si cela est obligatoire (par exemple pour le PMSI) ou s'il donne son consentement à travers l'utilisation de sa carte (pour une étude de nature médico-commerciale, par exemple) ;
- l'identifiant anonyme $IDA_{pat|Proj}$ n'utilise aucun secret dont la divulgation porterait atteinte à la vie privée des autres personnes (contrairement à l'utilisation d'une clé secrète commune pour tous les patients). De plus, puisque le calcul de l'empreinte $IDA_{pat|Proj}$ s'effectue au niveau de la carte, ID_{pat} reste toujours au sein de la carte ; il n'est jamais stocké isolement, et il n'est utilisé qu'afin de créer une entrée dans la base anonyme pour un projet donné (au niveau de l'hôpital) ;
- puisque ID_{proj} est spécifique à chaque projet, les risques de rapprochements non-autorisés des données de deux projets différents sont écartés, ou du moins sont peu vraisemblables ; de plus, les bases de données anonymes (par projet) sont isolées de l'extérieur de l'hôpital et sont soumises à des mesures strictes de contrôle d'accès ;
- sachant que l'empreinte $IDA_{pat|Proj}$ est toujours la même pour un patient et un projet donnés, il est possible que chaque projet puisse faire des rapprochements de données concernant un même patient.

Néanmoins, la transformation T1 ne permet pas de se prémunir contre certaines attaques où les intrus essayent de faire des rapprochements d'informations (concernant un projet donné) détenus par deux hôpitaux différents. En effet, supposant que le patient Paul a été traité à Rangueil et à Purpan, et que dans chacun de ces deux hôpitaux, Paul est consentant à l'utilisation de ses données pour un projet "*Proj□*".

---

[5] Même si la carte VITALE appartient à l'assuré et donc peut correspondre à plusieurs personnes (l'assuré et ses ayants-droits), l'identifiant ID_{pat} est individuel. Dès lors, sur une même carte peuvent figurer plusieurs identifiants correspondants à l'assuré et aux ayants-droit de moins de seize ans. Ceci sous entend un dispositif de classification de ces identifiants (par numéro d'ordre par exemple).

Supposons qu'un employé de Purpan, nommé Bob, sache que l'empreinte X (=$IDA_{Paul|Proj□}$) correspond à Paul. Supposons en plus, que Bob obtienne un accès à la base de donnée anonyme concernant *Proj□*, mais détenue par Rangueil. Dans ce cas, l'utilisateur malveillant Bob peut facilement établir le lien entre le patient Paul et ses données médicales (concernant *Proj□*) détenues par Rangueil.

Afin de faire face à ce type d'attaques, nous introduisons la *transformation asymétrique* **T2** au niveau de l'hôpital. Ainsi, avant de stocker les données dans les bases de données anonymes spécifiques à chaque projet, l'hôpital chiffre (par chiffrement asymétrique) l'identifiant $IDA_{pat|Proj}$ avec une clé $Ks_{hôp}$ spécifique à l'hôpital ; ("{}K" désigne un chiffrement avec K)□:

$$\textbf{(T2)} \qquad IDAhôp(pat|Proj) = \{IDApat|Proj\}\ Ks_{hôp}$$

Si on reprend le scénario précédent, l'utilisateur malveillant Bob ne peut guère croiser les identifiants anonymes des personnes car il ne dispose pas de la clé de déchiffrement Kp_{Purpan}. En effet, chaque hôpital détient sa clé $Ks_{hôp}$, tandis que la clé $Kp_{hôp}$ correspondante n'est détenue que par les projets.

Il est clair que les deux transformations (T1 et T2) effectuées au niveau des hôpitaux permettent d'avoir une grande robustesse vis-à-vis d'attaques ayant pour but de lever l'anonymat (ou de faire des rapprochements) de façon non autorisée. Pour autant, la procédure proposée reste assez flexible. En effet, si deux hôpitaux ($hôp_a$ et $hôp_b$) décident de fusionner un jour, il est tout à fait possible de relier les données concernant chaque patient□; que ces données proviennent de $hôp_a$ ou de $hôp_b$.

En effet, il suffit que chaque hôpital déchiffre ses données avec sa clé "$Kp_{hôp}$", puis chiffre le résultat avec la clé privée $Ks_{hôp_{ab}}$ du nouvel hôpital. Ainsi, si $IDA_{hôp_a(pat|Proj)}$ (respectivement $IDA_{hôp_b(pat|Proj)}$) désigne un identifiant anonyme au sein de l'hôpital $hôp_a$ (respectivement $hôp_b$)□; "[]K" désignant le déchiffrement avec K□:

- Le traitement effectué sur les anciennes données de l'hôpital $hôp_a$ est :

$$\{ [IDAhôp_a(pat|Proj)]\ Kp_{hôpa} \}\ Ks_{hôpab}□;$$

- Le traitement effectué sur les anciennes données de l'hôpital $hôp_b$ est,

$$\{ [IDAhôp_b(pat|Proj)]\ Kp_{hôpb} \}\ Ks_{hôpab}□;$$

Remarquons que les identifiants anonymes obtenus après ces traitements seront les mêmes dans les deux établissements (pour chaque base de donnée anonyme associé à un certain projet).

Pour les utilisateurs internes aux établissements de soins, les mécanismes de contrôles d'accès doivent interdire tout accès non-autorisé, tandis que des mécanismes de détection et de tolérance aux intrusions doivent renforcer les autres mesures de sécurité.

Transformations à la réception par les centres de traitements

Les données contenues dans les bases de données anonymes (au niveau des hôpitaux) subissent des transformations qui dépendent de l'identifiant anonyme $IDA_{proj|pat}$ et de la clé $Ks_{hôp}$. Pour retrouver les données qui lui sont destinées, chaque centre de traitement (correspondant à un projet) déchiffre les données qui lui sont envoyées par la clé $Kp_{hôp}$ de l'hôpital transmetteur□:

$$[IDAhôp(pat|Proj)]\ Kp_{hôp}$$

$$\text{et d'après } \textbf{(T2)}, \quad = [\ \{IDApat|Proj\}\ Ks_{hôp}\ ]\ Kp_{hôp}$$

$$= IDApat|Proj$$

Le centre de traitement retrouve ainsi les informations suffisantes et nécessaires aux traitements qu'il effectue. Ces informations sont associées aux identifiants anonymes $IDA_{pat|Proj}$, ce qui permet à chaque projet de chaîner les données de chaque patient.

Transformation avant la distribution aux utilisateurs finaux

Avant leur distribution aux utilisateurs finaux (recherche scientifique, publications, Web, presse, …), et afin de respecter le plus possible le principe du moindre privilège, les informations transférées peuvent devoir

subir un traitement de filtrage ciblé pour chaque catégorie d'utilisateurs. Il peut, par exemple, s'agir d'une agrégation, d'un appauvrissement des données, etc.

Si de plus, l'objectif est d'interdire à deux (ou plusieurs) utilisateurs finaux de recouper les informations, il convient d'appliquer une autre anonymisation (MD5, par exemple) avec une clé secrète $K_{util|proj}$.

$$IDApat|util = H(IDApat|Proj \mid Kutil|proj)$$

En fait, selon le besoin, ce dernier cas peut correspondre à deux situations (et donc procédures) différentes :

- si le but est de permettre à l'utilisateur de faire des chaînages dans le temps (par projet), la clé $K_{util|proj}$ doit être stockée au niveau du centre de traitement, de façon à pouvoir la réutiliser, à chaque fois que celui-ci souhaite transmettre d'autres informations à cet utilisateur ;
- à l'inverse, si le centre souhaite empêcher le chaînage dans le temps par les utilisateurs, la clé est générée aléatoirement à chaque distribution.

4.2 Discussion

La solution que nous proposons garantit les caractéristiques suivantes :

- Le patient doit donner explicitement son consentement pour toute utilisation non-obligatoire, mais souhaitable, de ses données.
- Les identifiants (ID_{proj}, ID_{pat}, $IDA_{pat|Proj}$ et $IDA_{pat|util}$) utilisés dans les diverses transformations sont situés dans des endroits différents ; de même, les clés ($Ks_{hôp}$, $Kp_{hôp}$) sont détenues par des personnes différentes. En effet, ID_{proj} ne concerne qu'un projet parmi d'autres ; ID_{pat} est spécifique à un patient, et cette information n'est jamais accessible : elle est stockée dans la carte du patient et n'en sort jamais; $Ks_{hôp}$ est spécifique à l'hôpital ; et $IDA_{pat|util}$ n'est destinée qu'à un utilisateur (ou type d'utilisateur) d'un projet donné. Il est donc pratiquement impossible de pouvoir faire des désanonymisations non autorisées.
- La solution résiste aux attaques par dictionnaire et à tous les niveaux : établissements de soins, centres de traitements et utilisateurs finaux.
- La séquence d'anonymisation (anonymisation en cascade) que nous proposons à différents niveaux, combinée avec des mécanismes de contrôle d'accès, permet de garantir, en toute robustesse, l'exigence de non inversibilité ainsi que le principe du moindre privilège.
- Les identifiants anonymes générés étant spécifiques à un secteur particulier (projet, domaine d'activité, centre d'intérêt, branche professionnelle, établissement, etc.), il est possible d'adapter la solution à chaque secteur (par exemple lorsque le centre de traitement est le seul utilisateur) ;
- Il est possible de fusionner les données de deux (voire de plusieurs) établissements sans compromettre la flexibilité et la sécurité.
- La manière selon laquelle l'information est distribuée et utilisée par l'utilisateur final est importante. Notre solution peut être adaptée pour tenir compte de la finalité du traitement ;
- Si un utilisateur final (chercheur dans le domaine des maladies orphelines par exemple) découvre une information qui nécessiterait de remonter aux identités des patients, il doit d'abord renvoyer ses résultats aux hôpitaux participant au projet concerné (probablement via les projets). L'inversibilité peut se faire de l'une des deux manières suivantes :
    a. Si l'hôpital originaire de ces informations dispose encore des bases de données (ou fichiers) anonymes et identifiées (dossiers administratifs et dossiers médiacaux) permettant d'établir le lien entre les identifiants anonymes, les numéros locaux de séjours, et les données médicales de ses patients, le professionnel soignant en charge de ce patient dans cet hôpital peut identifier le patient et l'informer des résultats.
    b. Si pour des raisons juridiques ou sécuritaires, l'hôpital a supprimé ces bases de données (ou fichiers), ou si le patient se présente à un autre centre de soin participant au projet, le consentement du patient recueilli à travers sa carte VITALE, permet de calculer $IDA_{pat|Proj}$ = H(ID_{proj} | ID_{pat}) et $IDAhôp_{(pat|Proj)}$ = {$IDA_{pat|Proj}$} $Kshô$p et donc d'établir le lien entre le patient, ses

identifiants anonymes et ses informations médicales. Une comparaison entre l'identifiant anonyme du patient et la liste des signalements (envoyée par l'utilisateur final) permettrait de déclencher une alarme demandant au patient s'il souhaite consulter l'information transmise.

Par ailleurs, nous pensons que cette solution ne résout pas tous les problèmes, et nous suggérons de la compléter, selon le cas étudié, par une combinaison de solutions techniques et organisationnelles :

- l'accès aux données doit être parfaitement contrôlé. Une *politique de contrôle d'accès* doit être définie et mise en place pour que les données ne soient accessibles qu'aux seuls utilisateurs habilités ;
- la spécification du système d'information et de l'architecture du réseau doit obéir à une politique globale de sécurité, et donc doit être adaptée aux besoins ;
- La définition de la politique de sécurité doit inclure une analyse des risques d'abduction ;
- la constitution de sous-bases de données régionales ou thématiques doit être contrôlée ;
- Il convient d'utiliser (si cela est possible) des anonymisations thématiques, de sorte que même si un utilisateur parvient à casser l'anonymisation, les risques d'abduction soient limités à un thème donné ;
- Il faut séparer les données d'identité des renseignements proprement médicaux. Bien entendu ce mécanisme ne peut être appliqué que dans des contextes particuliers ;
- Il est parfois souhaitable de renforcer la surveillance des utilisations qui sont faites des données, notamment en définissant et en mettant en œuvre des outils de *détection d'intrusion* ; en particulier, ces outils doivent permettre de détecter les requêtes, voire les enchaînements de requêtes, ayant un but malveillant (inférence de données, abus de pouvoir, *etc.*) ;
- nous préconisons également l'utilisation d'autres techniques comme le brouillage ou le filtrage, de façon à ne pas répondre à des requêtes statistiques si l'information demandée est trop précise ; etc.

Conclusion

Cet article répond à l'un des soucis récents, mais majeur, engendré par les nouvelles technologies de l'information et la communication : le respect de la vie privée et la protection de l'individu, dans une dimension électronique qui devient désormais omniprésente. Dans ce cadre, il analyse le problème d'anonymisation dans le domaine médical, identifie et étudie un certain nombre de scénarios représentatifs, et présente une démarche d'analyse mettant en correspondance des fonctionnalités d'anonymisation avec les solutions d'anonymisation adéquates. Enfin, il propose des procédures génériques, flexibles et adaptées aux besoins, objectifs et exigences de sécurité de ce domaine.

Bien que cette solution préconise plusieurs anonymisations en cascades, les mécanismes cryptographiques sur lesquels elle repose sont très peu coûteux en temps et ressources de calcul. Cette solution a été implémentée en utilisant la technologie des cartes à puces Java. Dans la suite de ce travail, nous envisageons de poursuivre cette étude en étudiant la complexité et en adaptant la solution à d'autres domaines où la protection de la vie privée est importante, notamment, le e-gouvernement ou le commerce électronique.

Remerciements

Cette étude a été soutenue partiellement par le réseau national de recherche en télécommunications (RNRT) dans le cadre du projet MP6. Un prototype est implémentée dans le cadre du projet européen PRIME du 6ème programme-cadre européen IST.

Références

[AFNOR 1997] AFNOR (1997). AFNOR, document de normalisation française, Fascicule de Documentation FD S 97-560.

[Castano et al. 1995] S. Castano, M. G. Fugini, G. Martella, P. Samarati, "*Database Security*", 1995, ACM press, ISBN: 0201593750, 456 pp.

[CC 1999] *Common Criteria for Information Technology Security Evaluation*, *Part* 1*: Introduction and general model*, 60 p., ISO/IEC 15408-1 (1999).

[CEN 1999] CEN/TC 251/WG I, *Norme prENV 13606-3: Health Informatics - Electronic Healthcare Record Communication*, n° 99-046, Comité Européen de Normalisation, 27 mai 1999.

[Circulaire 1989] Circulaire DH/PMSI n° 303 du 24 juillet 1989 relative à la généralisation du Programme de médicalisation (BOMS n° 89/46), Ministère de l'emploi et de la solidarité, France.

[Circulaire 1998] Circulaire n° 153 du 9 mars 1998 relative à la généralisation dans les établissements de santé sous dotation globale et ayant une activité de soins de suite ou de réadaptation d'un recueil de RHS, ministère de l'emploi et de la solidarité, France.

[Cuppens 2003] F. Cuppens, "Sécurité des bases de données", in *Sécurité des réseaux et des systèmes répartis*, (Y. Deswarte & L. Mé, eds), Traité IC2, Hermès, ISBN : 02-7462-0770-2, 264 pp, octobre 2003.

[Denning 1979] D. Denning et P. Denning, "Data Security". *ACM Computer Survey*, vol. 11, n° 3, septembre 1979, ACM Press, ISBN : 0360-0300, pp. 227-249.

[Directive 2002] Directive 2002/58/EC of the European Parliament on: *"the processing of personal data and the protection of privacy in the electronic communications sector"*; July 12, 2002; Official Journal L 201, 31-7-2002, p. 37-47.

[Directive 1995] Directive 95/46/CE of the European Parliament and the Council of the European union: "*On the protection of individuals";* October 24, 1995.

[Fabre *et al*. 1996] Jean-Charles Fabre, Yves Deswarte, Laurent Blain, «Tolérance aux fautes et sécurité par fragmentation-redondance-dissémination», *Technique et Science Informatiques (TSI)*, Vol.15(4), pp.405-427, 1996.

[Jeanneret *et al*. 2001] J.P. Jeanneret, D. Olivier, J. Chiffelle, "How to Protect Patient's Rigottes to Médical Secret in Official statistic", *Information Security Solutions Europe Conference* (*ISSE*), London, UK, 26-28 Septembre 2001.

[Loi 1978] Loi n°78-17 du 6 janvier 1978 relative à l'Informatique, aux fichiers et aux libertés, et décret d'application 78-774 du 17 juillet 1978.

[Loi 1991] Loi n° 91-748 du 31 juillet 1991 portant réforme hospitalière et décret n° 92-329 du 30 mars 1992 relatif au dossier médical et à l'information des personnes accueillies dans les établissements de santé publics et privés.

[Loi 1994] Loi 94-43 du 18 janvier 1994 relative à la santé publique et à la protection sociale, article 8.

[Ordonnance 1996] Ordonnance n° 96-346 du 24 avril 1996 portant réforme de "l'hospitalisation publique et privée des systèmes d'information et à l'organisation médicale dans les hôpitaux publics".

[Quantin *et al*. 1998] C. Quantin, H. Bouzelat, FA. Allaert, AM. Benhamiche, J. Faivre et L. Dusserre, "How to ensure data security of an epidemiological follow-up: quality assessment of an anonymous record linkage procedure", *International Journal of Medical Informatics* 49 (1998) 117-122.

[Résolution 1990] Resolution A/RES/45/95 of the General assembly of United Nations: *"Guidelines for the Regulation of Computerized Data Files"*; 14 December 1990.

[Recommandation 1997] Recommendations R(97)5 of the Council of Europe, *On The Protection of Medical Data Banks*, Council of Europe, Strasbourg, 13 February 1997.

[Shamir 1979] Shamir A., "How to Share a Secret", *Communications of the ACM*, Vol.22, n°11, November 1979, pp. 612-613.

[Trouessin 2001] Trouessin, G. (2001). "Sécurité et intimité des données à caractère personnel". *La Lettre d'ADELI n°42*, juillet 2001.