Une démarche méthodologique pour l'anonymisation de données personnelles sensibles

Anas Abou El Kalam[1], Yves Deswarte[1], Gilles Trouessin[2], and Emmanuel Cordonnier[3]

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

Résumé Cet article présente une nouvelle technique d'anonymisation destinée aux applications où les données personnelles doivent être cachées. Le problème de la protection de la vie privée ainsi que la base terminologique nécessaire à la compréhension de cet article sont d'abord présentés. Une démarche rigoureuse d'analyse des besoins et de choix de solutions est ensuite proposée. En particulier, seront expliqués les différences et les liens entre les besoins, les objectifs ainsi que les exigences d'anonymisation. Une caractérisation des solutions à choisir, à construire ou à mettre en ?uvre sera également proposée. L'analyse nous conduira à montrer l'intérêt majeur de l'utilisation des cartes à puces pour satisfaire les besoins de protection de la vie privée ; en particulier pour garder le secret de l'identifiant (ou des variables identifiantes en général) de l'utilisateur et pour exécuter la partie critique de procédure d'anonymisation. Ainsi, en fournissant sa carte, le citoyen (e.g., le patient dans le domaine médical) donne son consentement pour exploiter ses données anonymisées ; et pour chaque utilisation, un nouvel identifiant anonyme est générée à l'intérieur de la carte. Par ailleurs, cet article montre comment peut-on lever l'anonymat en respectant certains principes de bases, notamment le consentement du patient (en fournissant sa carte). Cet article montre que la technologie des cartes à puces peut jouer un rôle important pour la protection de la vie privée, ne serait ce que pour stocker un secret (un identifiant anonyme généré localement et aléatoirement à l'intérieur de la carte), et pour garder les procédures d'anonymisation (et donc l'utilisation des données anonymes), et de désanonymisation sous le contrôle permanent de l'utilisateur.

1 Problématique

Les applications informatiques émergentes utilisent des données dont le contenu est souvent sensible : réseaux de soins, recensements démographiques, commerce électronique, vote électronique, etc. D'une part, l'instauration des réseaux facilite le partage et la communication d'informations entre différentes structures ; d'autres part, elle pose des problèmes concrets de respect de la vie privée.

Conscients de ce problème, les législateurs imposent des exigences de confidentialité et de sécurité sur les données personnelles, en particulier par les réglementations internationales [1], européennes [2,3,4] et nationales [5,6]. Pour respecter cette législation, les systèmes (utilisant des données personnelles) ont besoin de mettre en ?uvre des méthodes et des moyens efficaces pour fournir (et justifier) un niveau élevé de sécurité. Mais malheureusement, la protection de la vie privée est souvent négligée ou mal étudiée. En l'occurrence, l'étude analytique des exigences d'anonymisation nous semble peu présente dans la majorité des solutions d'anonymisation actuelles ; plutôt que de se baser sur une méthodologie systématique, ces solutions sont souvent développées empiriquement.

Dans un souci de mener une réflexion bien fondée autour de ce problème plus que jamais d'actualité et d'y apporter des solutions réelles, cet article commence par définir le glossaire des termes élémentaires les plus couramment utilisés, associé au thème "anonymisation" (section suivante). Il propose ensuite une approche méthodologique pour préserver la vie privée, notamment en matière d'anonymisation des identifiants de personnes physiques figurant dans des fichiers informatisés (troisième section).

À cet égard, la démarche progressive présentée est essentiellement fondée sur l'identification des besoins, des objectifs et des exigences de sécurité, avant de définir ou de choisir la solution la plus adaptée à chaque problème lié au respect de la vie privée. Compte tenu des besoins de sécurité de l'application traitée, et en prenant en compte le contexte et la finalité des traitements que subiront les informations, nous identifions les premières questions cruciales à se poser :

- a-t-on besoin de chaînabilité, observabilité, anomynisation ou pseudonymisation ?
- Quel type de mécanisme (anonymisation irréversible ou inversible, appauvrissement de données, brouillage...) est le plus approprié ?
- Quelle forme de chaînage utiliser ?
- Quelle robustesse doit-on avoir et vis-à-vis de quoi ?
- ....

Afin de montrer l'intérêt pratique de notre méthodologie et de faire apparaître toutes ses facettes, nous allons l'appliquer dans un domaine où les données personnelles traitées sont d'une grande sensibilité. La troisième section présente ainsi un ensemble de scénarios représentatifs et en analyse les risques, les attaques, les besoins, les objectifs ainsi que les exigences d'anonymisations.

La dernière section tient compte des besoins déjà identifiés, pour présenter une nouvelle solution générique de génération et de gestion de données personnelles anonymisées. Cette solution montre, selon le cas, les transformations à appliquer aux données personnelles depuis leur collecte (au niveau des hôpitaux par

exemple), en passant par les centres de traitement (les associations de personnes diabétiques ou les centres des études cliniques, par exemple) jusqu'aux utilisateurs finaux (recherche scientifique, publications, Web, presse, par exemple). Dans un souci de respecter les législations, en particulier les recommandations de la norme européenne [7], le traitement choisi dans notre solution (chiffrement, anonymisation, filtrage de données, ...) tient compte du rôle et de l'établissement de rattachement de l'utilisateur, ainsi que de la finalité de l'utilisation prétendue.

Outre la méthodologie et la procédure globale d'anonymisation présentées dans cet article, une originalité fondamentale de la solution que nous proposons réside dans l'utilisation de la technologie des cartes à puces. Dans l'état actuel des connaissances, une carte à puce constitue un moyen matériel fiable et très difficilement falsifiable[4]. Ainsi, nous suggérons que la partie la plus critique de la procédure d'anonymisation soit exécutée au niveau d'une carte à puce (la carte VITALE, par exemple) appartenant à l'utilisateur (le patient, dans ce cas). Cette procédure est basée sur un secret, l'identifiant anonyme unique et individuel de l'utilisateur, généré aléatoirement (au sein de la carte), détenu par l'utilisateur (sur sa carte) et qui n'est jamais transmis à l'extérieur de la carte. Le secret reste donc sous le contrôle de l'utilisateur. Sauf en cas d'obligation légale, les données personnelles ne peuvent figurer dans une certaine base de données (pour une étude épidémiologique, par exemple) que si l'utilisateur donne son consentement en fournissant sa carte. La transformation cryptographique d'anonymisation s'effectue au sein de la carte ; par conséquent, l'identifiant initial de l'utilisateur n'est jamais connu en dehors de la carte.

En outre, afin de trouver un juste milieu entre l'anonymat et le lever d'anonymat, la désanonymisation doit également être sous le contrôle de l'utilisateur. En ce lieu, le rôle que joue la carte à puce, détenue par l'utilisateur, est incontestable : la désanonymisation n'est possible que si l'utilisateur donne son consentement en fournissant sa carte pour cette opération (la désanonymisation).

La fin de l'article présente une discussion qui montre que la solution proposée assure la sécurité (robustesse aux attaques par dictionnaires, par exemple) sans compromettre la flexibilité (supporter certains changements organisationnels comme le fusionnement de plusieurs établissements).

Après avoir exposé la problématique, et avant de développer davantage notre contribution, nous tenons à préciser les points suivants :

- même si l'étude de cas concerne le domaine santé-social, les préoccupations liées à l'anonymisation ne sont ni dédiées ni spécifiques à ce secteur ; la méthodologie ainsi que les solutions proposées restent applicables à une large gamme d'applications telles que celles citées au début de cet article ;
- même si nous abordons d'autres moyens (dispositifs) techniques et organisationnelles (comme le contrôle d'accès ou la détection d'intrusion) pouvant compléter notre solution, nous n'allons pas les détailler davantage ; ces dis-

[4] La falsification demeure très peu probable compte tenu des moyens à mettre en uvre pour réussir lattaque et des résultats obtenus même après lintrusion. La carte peut éventuellement être dotée dun petit programme enregistrant les tentatives daccès illicites à la carte.

positifs restent orthogonaux, et en les citant, notre but est tout simplement de montrer un cadre global ou notre proposition peut être intégrée.

2 Définitions

Notre analyse du domaine de la protection de la vie privée nous a permis de constater qu'il existe plusieurs nuances entre les termes utilisés par les différentes communautés scientifiques. À cet égard, citons par exemple la différence entre les législations européennes [2,3,4] qui visent à protéger les données personnelles et les lois Françaises [5,6] qui concernent les données nominatives. Dans notre vision, une donnée peut être "non-nominative" tout en restant "personnelle". En effet, comme il est parfois possible de ré-identifier des données anonymisées, ces données (anonymisées) peuvent perdre ou non leur caractère anonyme ; de même que des données nominatives peuvent perdre leur caractère nominatif. Dans tous les cas, nominatives, anonymes ou anonymisées, ces données restent à caractère personnel. Les anglo-saxons parlent d'ailleurs de "de-identification" au lieu d'anonymisation.

Afin d'éviter toute confusion et de faciliter la compréhension du reste de cet article, commençons tout d'abord par définir la base terminologique associé au thème de l'anonymisation des identifiants de personnes physiques figurant dans les fichiers informatisés.

Un identifiant d'une personne physique peut être défini comme une étiquette de nommage associée grâce à un système ou par une procédure d'identification à toute personne figurant dans une population donnée. Les deux propriétés essentielles à garantir sont l'atomicité ou la fiabilité :

- l'atomicité de l'identifiant est sa capacité à conserver ou à perdre la granularité élémentaire de l'information associée à une personne physique ;
- la fiabilité de l'identifiant est sa capacité à protéger de toutes formes possibles d'ambiguïtés inhérentes au système d'informations ;

Notons que même avec un identifiant atomique, il peut y avoir des doublons d'identifiant pour un même individu (comme des synonymes) ou à, l'inverse, des collisions d'identifiants d'individus distincts (comme des homonymes) : c'est la qualité de l'information amont à l'anonymisation qui présagera en grande majorité la qualité, et donc la fiabilité, du système d'anonymisation.

Par ailleurs, un identifiant peut revêtir deux formes : nominatif ou anonyme. L'identifiant est qualifié de nominatif s'il permet de déterminer sans ambiguïté la personne concernée ; dans le cas contraire, il est considéré comme anonyme.

Les critères communs (en anglais " *Common Critéria for Information Security Evaluation* ") qui sont maintenant une norme internationale [8] détaillent plusieurs classes fonctionnelles liées à la confidentialité, notamment l'anonymat, la pseudonymat, la chaînabilité et l'observabilité.

- L'anonymat peut être définie comme la propriété garantissant qu'un utilisateur peut utiliser une ressource ou un service sans révéler son identité ; autrement dit, l'impossibilité (pour d'autres utilisateurs) de déterminer le véritable nom de cet utilisateur.

- Le pseudonymat ajoute à l'anonymat le fait que l'utilisateur peut être tenu responsable de ses actes ; par exemple, en cas de litige ou d'enquÍte (lutte contre le blanchiment d'argent sale, par exemple), la propriété requise est la pseudonymat (plutôt que l'anonymat) car certaines informations personnelles doivent pouvoir être fournies aux autorités judiciaires.
- La non-chaînabilité représente l'impossibilité (pour d'autres utilisateurs) d'établir un lien entre différentes opérations réalisées par un même utilisateur ; par exemple, en interdisant la fusion ou le croisement d'informations à partir de différents fichiers ou bases de données.
- La non-observabilité garantit qu'un utilisateur peut utiliser une ressource ou un service sans que d'autres utilisateurs soient capables de déterminer si une opération (l'utilisation d'une ressource ou d'un service) est en cours.

3 Démarche d'analyse

L'analyse des solutions d'anonymisation utilisées dans les systèmes de santé des pays européens [9] nous a permis de détecter certaines défaillances, souvent liées à une absence d'une démarche analytique préalable des risques, des besoins, des exigences ainsi que des objectifs de sécurité. Comme pour beaucoup de solutions d'ordre sécurité, nous pensons que l'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.

Une fonctionnalité d'anonymisation peut être exprimée selon un des trois niveaux d'attente suivants classés par ordre croissant de force :

- le besoin d'anonymisation, représente les attentes de l'utilisateur ; généralement sous une forme qui n'est pas toujours très bien explicité ni aisé à formaliser ;
- l'objectif d'anonymisation, spécifie le niveau de sécurité à atteindre ou les menaces à éviter (comment satisfaire les exigences ?) ;
- l'exigence d'anonymisation, représente la façon d'exprimer le besoin ; dans la mesure du possible, très proche d'un formalisme clair et d'une sémantique non-ambiguë.

3.1 Besoins d'anonymisation

Les besoins de protection de la vie privée peuvent être d'ordre général, comme ceux identifiés dans la première section (problématique), mais aussi et surtout spécifiques au système étudié. Dans les systèmes de santé par exemple, une liste non exhaustive des besoins d'anonymisation pourrait être :

- outre le nom, le prénom et le numéro de sécurité social, les données les plus sensibles sont la date de naissance (parfois, seulement l'année de naissance est nécessaire), l'adresse (parfois, seulement la région est intéressante à connaître), la nationalité (il est probablement plus judicieux dans certains cas d'effectuer des regroupements)....

- Les données personnelles à cacher correspondent non seulement aux données directement nominatives (comme le nom, le prénom, le numéro de sécurité sociale, le sexe et l'adresse), mais aussi aux données indirectement nominatives (qui caractérisent la personne). En effet, 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).
- Conformément à la législation en vigueur, certaines données collectées doivent être détruites après une durée limitée. Actuellement le règlement des archives hospitalières, impose des délais de conservation de 70 ans pour les dossiers de pédiatrie, de neurologie, de stomatologie et de maladies chroniques.
- Les données d'un patient n'apparaissent dans une certaine base de données (pour une étude médico-commerciale, par exemple) que si c'est obligatoire ou si ce patient donne son consentement ; de la même manière, il nous paraît évident d'avoir le consentement du patient pour lever l'anonymat.
- L'anonymisation est fondée sur un secret, l'identifiant anonyme du patient, qui n'est utilisé que pour chaîner les données médicales du patient, tout en respectant sa vie privée.

Certains des besoins identifiés montrent que les cartes à puces peuvent être une alternative innovante contribuant à la protection de la vie privée. En effet :

- avec la technologie actuelle des cartes à puces, il est possible d'anonymiser toutes les données identifiées comme personnelles[5].
- L'utilisation de la carte est le moyen idéal pour matérialiser le consentement du patient, du moins implicitement.
- La carte est un moyen supposé fiable pour protéger un secret (l'identifiant anonyme du patient) contre toute attaque visant sa lecture ou sa modification illicites. Si en plus, la transformation cryptographique (l'anonymisation ou la désanonymisation) est exécutée au sein de la carte, on peut garantir que le secret (l'identifiant) n'est jamais transmis (connu) en dehors de la carte.

La section suivante commence par identifier ce que nous entendons par objectifs et exigences d'anonymisation. La fin de la section 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. La dernière section de cet article achève l'application de notre méthodologie en proposant une nouvelle procédure d'anonymisation.

[5] Les cartes à puces actuelles sont capables de stocker les identifiants et supportent la réalisation de traitements simples comme MD5 ou SHA (pour lanonymisation).

3.2 Objectifs d'anonymisation

Un objectif d'anonymisation est défini en fonction de l'une des trois propriétés suivantes applicables à la fonction d'anonymisation [10] :

- *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[6], 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 surveillance 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 [8] (cf. section 2).

3.3 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 données 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 1, 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") [11].

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

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

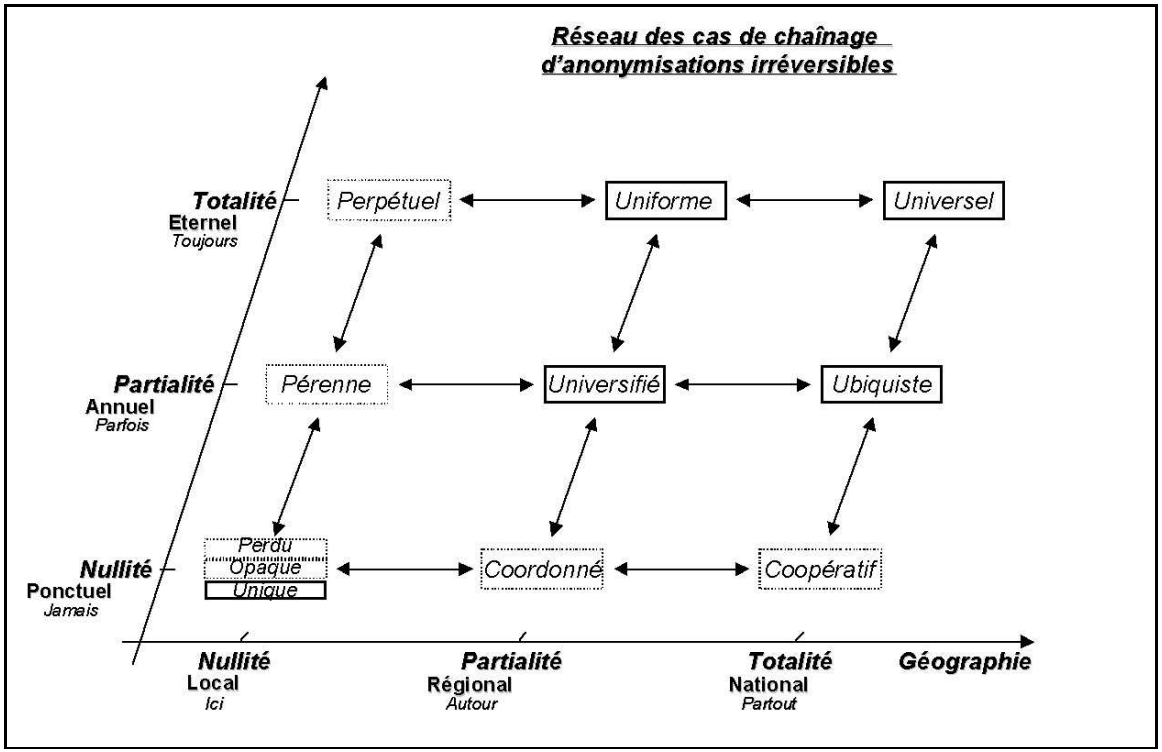

Fig. 1. Anonymisation en cascade : de l'universalité jusqu'à l'unicité

dans les quelques jours qui suivent, fait un test de dosage, alors le résultat du dépistage était positif ;
- *inductive* : s'apparente souvent à des raisonnements de type loi des grands nombres sans forcément l'appliquer sur de grandes échelles ; cela consiste par exemple, à induire qu'un tel 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 ;
- *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.

3.4 Choix de solutions

Compte tenu du système, les sections précédentes tentent de donner sens aux bonnes questions à se poser :
- le type de réversibilité : anonymisation irréversible, réversible ou bien inversible ;
- le type de chaînage : chaînage spatial, temporel ou bien spatio-temporel ?
- la forme du chaînage : chaînage nul, partiel ou total ?
- la robustesse à la réversion : réversion directe (inversion) ou bien à indirecte (reconstruction) ;
- la robustesse à l'inférence : inférence déductive, inductive, abductive ou bien adductive ?
- ...

En réponse aux questions posées supra, il est possible de définir une politique d'anonymisation, au travers des choix suivants qui parfois dérivent naturellement en fonction des fonctionnalités quand elles sont correctement exprimées ou bien restent ouvertes s'il n'y a pas de recommandation :
- type de solution : organisationnelle, mécanisme cryptographique ou fonction à sens unique ?
- pluralité de la solution : mono-anonymisation, bi-anonymisation ou bien multi-anonymisation ?
- interopérabilité de la solution : par transcodage (manuel), translation (mathématique) ou bien transformation (automatique) ?
- ...

4 Analyse de scénarios du domaine santé-social

4.1 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 2 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.

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

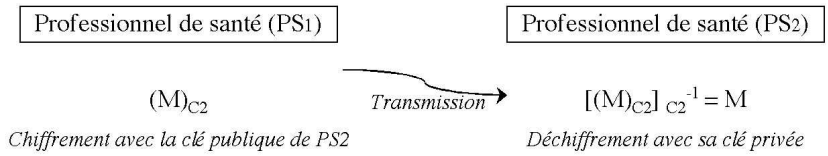

Fig. 2. Échange de données chiffrées entre professionnels de santé

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

4.2 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. En effet, l'article L4134-4 du code de la santé publique ainsi que l'article 81 de la loi 94-43 [14] précisent que " les médecins conventionnés exerçant à titre libéral dans la circonscription de l'union sont tenus de faire parvenir à l'union les informations mentionnées à l'article L.161-29 du code de la sécurité sociale relatives à leur activité, sans que ces informations puissent être nominatives à l'égard des assurés sociaux ou de leurs ayants-droit ou, à défaut, à condition qu'elles ne comportent ni leur nom, ni leur prénom, ni leur numéro d'inscription au Répertoire national d'identification des personnes physiques. Ces informations ne sont pas nominatives à l'égard des médecins ". Ces textes ajoutent que " L'anonymat (des médecins) ne peut être levé qu'afin d'analyser les résultats d'études menées dans le cadre de l'évaluation

des comportements et des pratiques professionnelles en vue de la qualité des soins ". Le scénario décrit dans cette section est résumé dans la figure 3. En tenant compte des besoins exprimés à travers cette législation, il conviendrait d'utiliser des pseudonymes pour les médecins et pour les patients. Les objectifs sont alors :

- 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[7]) ;
- 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.

4.3 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 [13] pour l'activité de MCO (Médecine, Chirurgie, Obstétrique). Son utilisation à des fins budgétaires a été formalisée par la circulaire du 7 décembre 1996 [14]. Il a été étendu aux établissements privés par les ordonnances du 24 avril 1996. La circulaire du 9 mars 1998 [15] 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 [16], le décret du 27/07/94 [17] ainsi que les arrêtés des 20/09/1994, 22/07/1996 et 29/07/1998 [18].

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

[7] Le principe du moindre privilège impose que tout utilisateur ne doit pouvoir accéder à un instant donné quaux informations et services strictement nécessaires pour laccomplissement du travail qui lui a été confié.

de naissance, nom et prénom, par exemple) 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 donnée. À partir des RUM récupérés et validés, 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 4.

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 anonymisation (au sens des critères communs [8]) ;
- 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.).

4.4 Traitement des 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 dispensent 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 : ne paraît pas nécessaire de disposer de la profession précise ; une simple mention des catégories socio-professionnelles selon la nomenclature de l'INSEE 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.

Actuellement en France, il semble que le but est d'assurer un suivi des cas et de mesurer l'évolution du SIDA. Il est donc nécessaire de collecter des données individualisées afin de permettre de détecter les doublons, de vérifier les données auprès des professionnels de santé, etc. En ce qui concerne l'appauvrissement des données, l'Institut de Veille Sanitaire a montré que l'utilisation d'informations indirectement nominatives figurant sur les déclarations du SIDA (initiales du nom et prénom, date de naissance, département de domicile) permet de repérer plus de 99% des doublons. Un appauvrissement trop important des données peut donc fausser les statistiques et remettre en cause la fiabilité scientifique de la surveillance épidémiologique.

4.5 Traitement des données statistiques

En aucun cas, les données médicales à caractère personnel ne peuvent être manipulées pour des traitements à des fins non-épidémiologiques, en l'occurrence, 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 [19,20,21]. La référence [22] explique que la sécurité dans les bases de données statistiques est un problème réel, que plusieurs suggestions apparaissent dans la littérature, mais qu'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 données 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.

4.6 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...

Si, par exemple, ces études mettent 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 $T_{\text{recouvrement}}$. 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 (pseudonymisation dans le sens des critères communs) : seules des personnes habilitées peuvent lever l'anonymat (*médecins conseils, médecins inspecteurs, médecins traitants*) et seulement quand c'est nécessaire (principe du moindre privilège).

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 recoupe-

ments entre plusieurs données déjà recueillies, et les compléter a posteriori. Là encore, la pseudonymisation semble nécessaire.

5 Une nouvelle solution générique

5.1 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 et met en oeuvre les recommandations de la norme européenne [7].

Pour cela, et comme indiqué sur la figure 5 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).

Transformations au niveau des fournisseurs de données sensibles 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,
- des bases de données anonymes, dont chacune contient les informations nécessaires et suffisantes pour un projet donné (un centre de traitement).

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 "$\mathrm{IDA}_{\mathrm{pat—Proj}}$", un identifiant anonyme par personne et par projet, à partir des deux identifiants :

- "$\mathrm{ID}_{\mathrm{proj}}$", l'identifiant du projet, qui est détenu par les établissements de soins (hôpitaux, cliniques) ;
- "$\mathrm{ID}_{\mathrm{pat}}$", l'identifiant anonyme détenu par le patient sur la carte VITALE (rappelons que $\mathrm{ID}_{\mathrm{pat}}$ 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 $\mathrm{ID}_{\mathrm{proj}}$ (l'identifiant du projet concerné par la base de donnée) à la carte ; celle-ci contient déjà $\mathrm{ID}_{\mathrm{pat}}$ (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 $(\mathrm{ID}_{\mathrm{proj}}|\mathrm{ID}_{\mathrm{pat}})$, la concaténation de $\mathrm{ID}_{\mathrm{proj}}$ et $\mathrm{ID}_{\mathrm{pat}}$:

$$\text{T1} \qquad \mathrm{IDA}_{\mathrm{pat—proj}} = H(\mathrm{ID}_{\mathrm{proj}}|\mathrm{ID}_{\mathrm{pat}})$$

La transformation T1, réalisée au sein de la carte VITALE[8] du patient, et produisant l'empreinte $H(\mathrm{ID}_{\mathrm{proj}}|\mathrm{ID}_{\mathrm{pat}})$, 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 la fourniture de son identifiant (pour une étude de nature médico-commerciale, par exemple) ;
- l'identifiant anonyme $\mathrm{IDA}_{\mathrm{pat—proj}}$ 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 $\mathrm{IDA}_{\mathrm{pat—proj}}$ s'effectue au niveau de la carte, $\mathrm{ID}_{\mathrm{pat}}$ 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 $\mathrm{ID}_{\mathrm{proj}}$ 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 $\mathrm{IDA}_{\mathrm{pat—proj}}$ 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 de l'utilisation de ses données pour un projet "Projα". Supposant qu'un employé de Purpan, nommé Bob, sait que l'empreinte X (= $\mathrm{IDA}_{\mathrm{Paul|Proj}_\alpha}$) correspond à Paul. Supposant en plus, que Bob arrive à s'emparer de la base de donnée anonyme concernant Projα, 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 Projα) détenues par Rangueil (en plus de celles détenues par Purpan, puisque Bob travaille à Purpan).

[8] Il sagira probablement dune nouvelle génération de cartes VITALE qui supporteraient la réalisation de traitements simples comme MD5 ou SHA.

Afin de faire face à ce type d'attaques, nous introduisant 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 (chiffrement asymétrique) l'identifiant $\mathrm{IDA}_{\mathrm{pat—proj}}$ avec une clé $Ks_{\mathrm{hôp}}$ spécifique à l'hôpital ; ("$\{\}_K$" désigne un chiffrement avec K) :

$$\text{T2} \qquad \mathrm{IDA}_{\mathrm{hôp}}(\mathrm{pat—Proj}) = \{IDA\}_{Ks_{\mathrm{hôp}}}$$

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

Il est facile de constater 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 ($\mathrm{hôp}_a$ et $\mathrm{hôp}_b$) décident de fusionner un jour, il est tout à fait possible de relier les données concernant chaque patient ; que ces données proviennent de $\mathrm{hôp}_a$ ou $\mathrm{hôp}_b$.

En effet, il suffit que chaque hôpital déchiffre ses données avec sa clé $Kp_{\mathrm{hôp}}$, puis chiffre le résultat avec la clé privée $Kp_{\mathrm{hôp}_{ab}}$ du nouvel hôpital. Ainsi, si $\mathrm{IDA}_{\mathrm{hôp}_a}(pat|Proj)$ (respectivement $\mathrm{IDA}_{\mathrm{hôp}_b}(pat|Proj)$) désigne un identifiant anonyme au sein de l'hôpital $\mathrm{hôp}_a$ (respectivement $\mathrm{hôp}_b$) ; $[]_K$ désignant le déchiffrement avec K :

– Le traitement effectué sur les anciennes données de l'hôpital $\mathrm{hôp}_a$ est :

$$\{[\mathrm{IDA}_{\mathrm{hôp}_a}(\mathrm{pat|Proj})]Kp_{\mathrm{hôp}_a}\}Ks_{\mathrm{hôp}_{ab}}$$

– Le traitement effectué sur les anciennes données de l'hôpital $\mathrm{hôp}_b$ est,

$$\{[\mathrm{IDA}_{\mathrm{hôp}_b}(\mathrm{pat|Proj})]Kp_{\mathrm{hôp}_b}\}Ks_{\mathrm{hôp}_{ab}}$$

Remarquons que les codes de liaisons obtenus 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 en amont des 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 $\mathrm{IDA}_{\mathrm{pat—proj}}$ et de la clé $Ks_{\mathrm{hôp}}$. 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 $Kp_{\text{hôp}}$ de l'hôpital transmetteur (d'après (T2)) :

$$[\text{IDA}_{\text{hôp}(pat|Proj)}]Kp_{\text{hôp}} = [\{\text{IDA}_{\text{pat—proj}}\}Ks_{\text{hôp}}]Kp_{\text{hôp}} = \text{IDA}_{\text{pat—proj}}$$

Le centre de traitement retrouve ainsi les informations suffisantes et nécessaires aux traitements qu'il effectue. Ces informations sont associées aux identifiants anonymes $\text{IDA}_{\text{pat—proj}}$, 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 éventuellement 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 de sécurité 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 $K_{\text{util—proj}}$.

$$\text{IDA}_{pat|util} = H(\text{IDA}_{pat|Proj}|K_{util|proj})$$

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.

Désanonymisation L'analyse des scénarios (cf. Section 4) montre qu'il est parfois souhaitable, voir nécessaire de lever l'anonymat. En outre, l'étude des besoins préconise le consentement du patient pour la réalisation de cette procédure (désanonymisation). Afin de satisfaire ces besoins, nous proposons que 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). En se présentant à un de ces hôpitaux, et en fournissant sa carte VITALE, le patient donne son consentement pour lever l'anonymat, et associer ainsi les nouvelles informations (résultat de la recherche scientifique, par exemple) à l'identité réelle du patient.

Celui-ci pourrait ainsi bénéficier de ces résultats. ID_{pat} figurant sur la carte, ainsi que ID_{Proj} et $Ks_{\text{hôp}}$ fournis par le système de l'hôpital, permettraient de calculer

$$\text{IDA}_{pat|Proj} = H(\text{ID}_{proj|}\text{ID}_{pat})$$

et

$$\text{IDAhÙp}_{(}pat|Proj) = \{\text{IDA}_{pat|Proj}\}_{Ks_{\text{hôp}}}.$$

C'est donc la seule façon 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 inversions (envoyée par l'utilisateur final) permettrait de déclencher une alarme demandant au patient s'il souhaite consulter l'information transmise.

Discussion La solution que nous proposons vise à garantir les points suivants :

- L'identifiant anonyme du patient est un secret qui est protégé de tout accès illicite (en lecture ou en modification). Cette donnée sensible est générée aléatoirement au sein de la carte, dispositif supposé fiable et très difficilement falsifiable ; de plus, les identifiants anonymes spécifiques aux projets sont calculés au sein de la carte. Le secret (ID_{pat}) n'est donc jamais transmis en dehors de la carte, ni altéré illicitement.
- L'utilisateur doit donner explicitement son consentement pour toute utilisation non-obligatoire, mais souhaitable, de ses données. De cette manière, tout chaînage de données personnelles ainsi que toute procédure destinée à lever l'anonymat, sont strictement contrôlés par l'utilisateur. La solution résiste aux attaques par dictionnaire et à tous les niveaux : établissements fournisseurs de données sensibles, centres de traitements et utilisateurs finaux.
- La séquence d'anonymisation (anonymisation en cascade) que nous proposons à différents niveaux, combinée avec des mécanismes de contrôles 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 (voir 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 adoptée pour tenir compte de la finalité du traitement.

Actuellement, les hôpitaux français utilisent l'algorithme de hachage SHA (Standard Hash Algorithm) pour transformer, d'une manière irréversible, les variables 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 [**?**]. Afin de chaîner les informations concernant le même patient, le code anonyme obtenu (après anonymisation) est toujours le même pour un individu donné.

deux clés ont été ajoutées à l'algorithme de hachage SHA. La première clé k_1, utilisée par tous les émetteurs des données (hôpitaux et médecins), est concaténée

à l'identité. Une fonction de hachage est ensuite appliquée au résultat :

$$\text{empreinte}_1 = H(k_1|\text{identité}).$$

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 (voir figure 6) :

$$\text{empreinte}_2 = H(k_2 + \text{empreinte}_1).$$

Il est évident que ce protocole s'avère complexe et risqué. En effet, il nécessite une distribution de la même clé secrète à tous les fournisseurs d'informations (médecins libéraux, hôpitaux, cliniques...), tout en supposant que cette clé doit rester secrète. Si une clé 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 (sachant que les empreintes dépendent de la clé supposée être toujours la même et chez tous les fournisseurs d'informations) ? Si ce problème survient, la seule solution envisageable consiste à appliquer une autre transformation à toute la base de données, solution qui n'est guère aisée.

À l'inverse, dans notre solution, les identifiants (ID_{pat}, ID_{Proj}, $\text{IDA}_{pat|Proj}$ et $\text{IDA}_{pat|Util}$) utilisés dans les diverses transformations sont situés dans des endroits différents, et les clés ($Ks_{\text{hôp}}$, $Kp_{\text{hôp}}$) sont détenues par des personnes différentes. En outre, ID_{Proj} est spécifique à un seul projet ; la paire de clés ($Ks_{\text{hôp}}$, $Kp_{\text{hôp}}$) est relié à un seul hôpital ; $\text{IDA}_{pat|Util}$ est destinée à un seul utilisateur ; etc. Il est donc pratiquement impossible de lever illicitement l'anonymat ou de réussir des inférences non autorisées. De plus, puisque ID_{pat} est spécifique à un seul patient (et ne figure que sur sa carte), sa divulgation (qui reste une tâche très difficile) ne compromet guère la sécurité totale du système.

Par ailleurs, nous pensons que la solution idéale n'existe pas, et nous suggérons de compléter notre solution, 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 oeuvre 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...) ;
- 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 ; ....

6 Conclusion

Cet article propose d'utiliser la technologie des cartes à puces pour répondre 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, 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 protection de la vie privée.

La solution proposée est en phase finale d'implémentation à travers un scénario complet du domaine médical, allant de l'enregistrement des données au niveau des hôpitaux jusqu'aux utilisations finales (recherche scientifique, Web, presse...).

Nous envisageons de poursuivre ce travail en étudiant la complexité et en adaptant la solution à d'autres exemples comme les recensements démographiques, le commerce électronique ou le vote électronique.

Références

1. Résolution A/RES/45/95 Assemblée générale des Nations Unies, Principes directeurs pour la réglementation des fichiers personnels informatisés, 14 Décembre 1990.
2. Directive 95/46/CE du Parlement Européen, adoptée par le Conseil Européen le 24 juillet 1995, On the protection of individuals with regard to the processing of personnal data and on the free movement of such data, 1995.
3. Directive du Parlement Européen n° 2002/58/EC concernant "le traitement des données à caractère personnel et la protection de la vie privée dans le secteur de

télécommunications électroniques", 12 juillet 2002, Journal Officiel L 201, 31-7-2002, pp. 37-47.

4. Recommandations du Conseil de l'Europe, R(97)5, On The Protection of Medical Data Banks, Council of Europe, Strasbourg, 13 février 1997.
5. Loi 78-17 du 6 janvier 1978 relative à l'Informatique, aux fichiers et aux libertés, Journal officiel de la République française, pp. 227-231, décret d'application 78-774 du 17 juillet 1978, pp. 2906-2907.
6. Loi 2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé, article L. 1111-7.
7. 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.
8. Common Criteria for Information Technology Security Evaluation, Part 1 : Introduction and general model, 60 p., ISO/IEC 15408-1 (1999).
9. A. Abou El Kalam, Y. Deswarte, G. Trouessin, E. Cordonnier, "Gestion des données médicales anonymisées : problèmes et solutions",2ème conférence francophone en gestion et ingénierie des systèmes hospitaliers (GISEH'04), 9-11 septembre 2004, Mons, Belgique, 2004, (à paraitre).
10. G. Trouessin, G. "L'évolution des normes de sécurité vers plus d'auditabilité des systèmes d'information". Colloque AIM à l'HEGP : « Présent et avenir des systèmes d'information et de communication hospitaliers », 23-24 mai 2002 Springer-Verlag.
11. AFNOR, document de normalisation française, Fascicule de Documentation FD S 97-560.
12. Loi 94-43 du 18 janvier 1994 relative à la santé publique et à la protection sociale, article 8.
13. 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.
14. 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".
15. 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.
16. Loi 91-748 du 31 juillet 1991 portant réforme hospitalière.
17. Décret n° 94-666 du 27 juillet 1994 relatif aux systèmes d'information médicale et l'analyse de l'activité des établissements de santé publics et privés sous compétence tarifaire de l'État, modifié par le décret n° 98-63 du 2 février 1998.
18. Arrêté du 29 juillet 1998 relatif au recueil et traitement des données d'activité médicale par les établissements de santé publics et privés financés par dotation globale visées à l'article L. 710-16-1 du même code et à la transmission aux agences régionales de l'hospitalisation et à l'État d'informations issues de ce traitement (JO, 26 août 1998).
19. D. Denning et P. Denning, "Data Security". ACM Computer Survey, vol. 11, n° 3, septembre 1979, ACM Press, ISBN : 0360-0300, pp. 227-249.

20. F. Cuppens, "Sécurité des bases de données", in Sécurité des réseaux et des systèmes répartis, (Yves Deswarte & Ludovic Mé, eds), Traité IC2, Hermès, ISBN : 02-7462-0770-2, 264 pp, octobre 2003.
21. A. Abou El Kalam, "Modèles et politiques de sécurité pour les domaines de la santé et des affaires sociales", Thèse de doctorat, Institut National Polytechnique de Toulouse, 183 pp., 04 décembre 2003 (Rapport LAAS 03578).
22. 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.

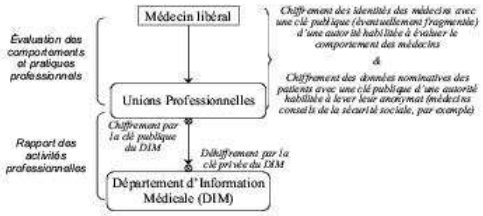

Fig. 3. Manipulations des identités au niveau des unions professionnelles

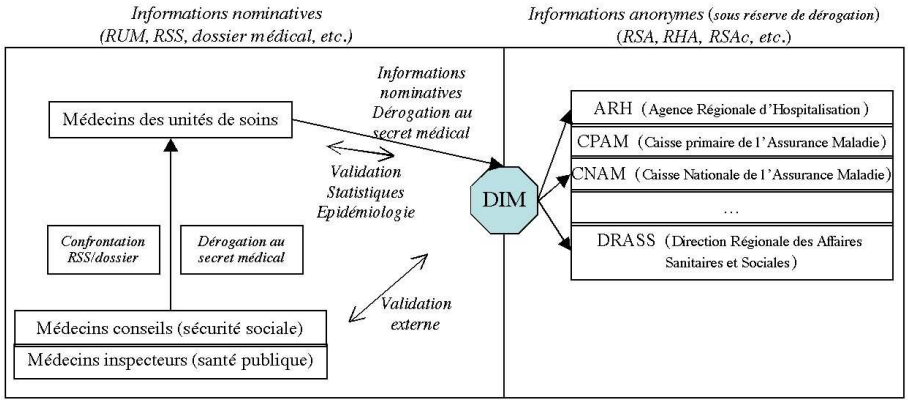

Fig. 4. Frontières des données nominatives, anonymes et anonymisables.

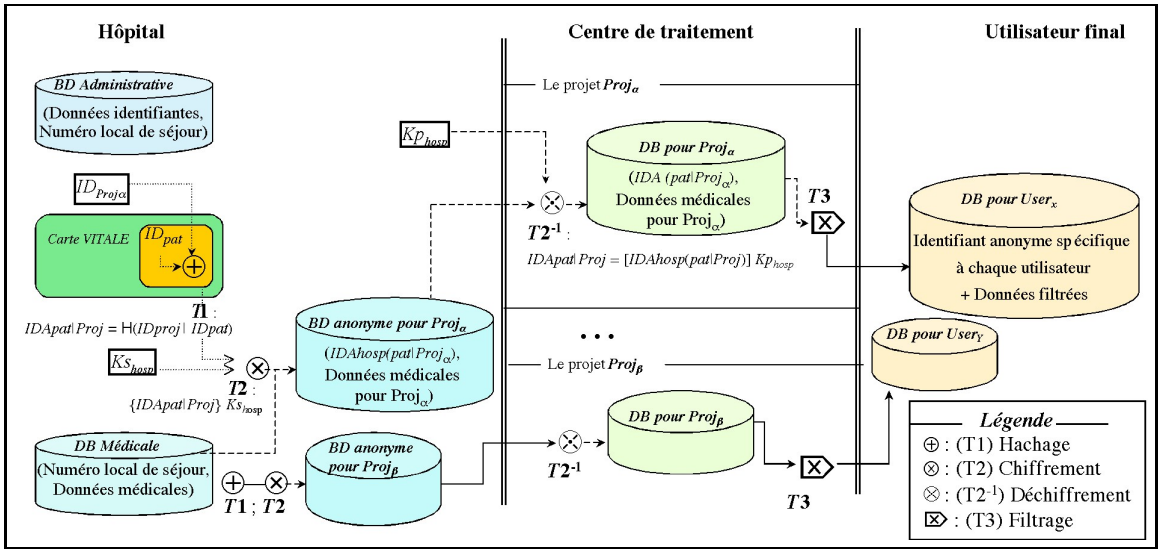

Fig. 5. Procédure d'anonymisation proposée.

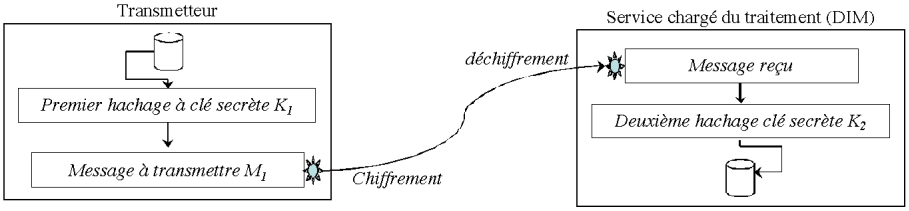

Fig. 6. Les grandes lignes de la procédure de hachage utilisée actuellement dans les hôpitaux français