SMARTCARD-BASED ANONYMIZATION

Anas Abou El Kalam, Yves Deswarte,
LAAS-CNRS, 7 avenue du colonel Roche, 31077 Toulouse Cedex 4; {Deswarte, anas}@laas.fr

Gilles Trouessin,
ERNST & 3YOUNG, 1 place Alfonse Jourdain, 31000 Toulouse; gilles.trouessin@fr.ey.com

Emmanuel Cordonnier
ETIAM. 20 Rue du Pr Jean Pecker, 35000 Rennes; emmanuel.cordonnier@etiam.com

Abstract: This paper presents a new technique for anonymizing personal data for studies in which the real name of the person has to be hidden. Firstly, the privacy problem is introduced and a set of related terminology is then presented. Then, we suggest a rigorous approach to define anonymization requirements, as well as how to characterize, select and build solutions. This analysis shows that the most important privacy needs can be met by using smartcards to carry out the critical part of the anonymizaton procedure. By supplying his card, the citizen (e.g., the patient in the medical field) gives his consent to exploit his anonymized data; and for each use, a new anonymous identifier is generated within the card. In the same way, reversing the anonymity is possible only if the patient presents his personal smartcard (which implies that he gives his consent). In this way, the use of the smartcard seems be the most suitable means of keeping the secret as well as the anonymization and the disanonymization procedures under the patient control.

Key words: Privacy, anonymization.

1. INTRODUCTION

Privacy is becoming a critical issue in many emerging applications in networked systems: E-commerce, E-government, E-health, etc. Of course, the networks development facilitates data communication and sharing, but such data are often personal and sensitive. Consequently, the citizen's privacy could be easily endangered, e.g., by inferring personal information.

International [1] and European legislations are more and more worried about protecting personal data and privacy is nowadays considered as a constitutional right in many countries [2, 3, 4]. To comply with these laws, systems dealing with personal data need more and more security and reliability. But unfortunately, the real privacy requirements are sometimes neglected, and solutions intended to solve some privacy problems are often developed empirically. What we first need is to have a systematic methodology to create and manage securely anonymized data.

The next section presents an analytic approach, based on risks analysis, security requirement identification, security objective definition, requirement formalization, and finally, solutions characterization. We have to reply questions such as: which data must be protected and for how long? Do we need unlinkability, unobservability anonymity, or pseudonimity? Where the cryptographic transformations should be carried out? Which transformation should be used, and do we need complementary technical or organizational measures? What is the best trade-off between robustness and flexibility? Etc.

Following our methodology, we suggest in Section 3 a new generic solution to anonymize and link identities. We will show that the use of a smart card is important to meet better privacy needs.

Let us take for example the healthcare systems (the reasoning remains the same for other applications such as E-commerce or E-government). It is clear that if a patient anonymous identifier is used to link the anonymized medical data coming from different sources or at different periods for the same patient, this identifier has to be kept secret to preserve the patient's privacy. In addition, the patient must rationally gives his consent when his data are anonimized or disanonimized. For all these reasons, and for other reasons that we will explain below, we believe that the use of smartcards could be a suitable solution. We suggest that the most critical part of the anonymization procedure should be carried out in a *personal medical data smart card.* The anonymization procedure should be based on a secret, the *patient anonymous identifier,* which is a randomly generated number, stored in the card, and never transmitted out of the card. This anonymous identifier is used to link uniquely the patient's anonymized medical data to the patient, while preserving his privacy.

In this way, the secret is held under the patient control. Except when it is legally mandatory, patient medical data may appear in a certain database only if, by supplying his card, he gives his consent to exploit his medical data as part of a certain project (e.g., epidemiologic research concerning a rare disease). In the same way, the patient consent is necessary to reverse the anonymity, for example, in order to refine epidemiological studies.

In the last section, we show that our solution presents a large flexibility and efficiency in that the data transformation procedures (e.g., anonymization, impoverishment) depend on the purpose of use (i.e., the reason why the data are collected). Moreover, our solution resists dictionary attacks, respects the least privilege principle and fulfills the European legislation requirements. It is also adaptable to non-medical areas, such as demographic studies, and supports (without compromising the security nor the flexibility) some organizational changes like the merging of hospitals.

2. ANALYTIC APPROACH

Based on previous works, we present a set of concepts that are necessary to cover this topic. *Anonymity* can be defined as the state of being not identifiable within a set of subjects. *Pseudonymity* add accountability to anonymity, more precisely, pseudonymity ensures that a user may use a resource or a service without disclosing its identity, but can still be made accountable for that use. *Unlinkability* between two or more items (operations, users initiating operations, terminals where operations were launched from, etc.) means that, within the system, it is not possible to distinguish if these items are related or not. *Unobservability* ensures that a user may use a resource or service without others, especially third parties, being able to observe that the resource or service is being used; therefore, the intent is not only to hide the identity of the person using a resource or a service, but even to hide if the resource or service is used [7; 9].

Besides, the privacy protection necessitates specifying two major categories of concepts:

the *request* (demand) in the form of needs to be satisfied;

the *response* in the form of security functionalities and solutions.

Like most of security functionalities, anonymization analysis can be expressed with regard to three levels of user expectations:

The *anonymization* needs represent the user expectations; generally, their form is neither very explicit nor very simple to formalize.

The *anonymization* objectives specify the information to protect, the threats (against privacy) to avoid, the security level to reach, etc.

The *anonymization* requirements represent how to express the needs (and the threats to counter) with a non-ambiguous semantics and, as far as possible, with a formal system.

The anonymization needs depend on the studied system. For instance in the healthcare systems, some anonymization needs could be:

1. both directly nominative information (identities, sex, addresses) and indirectly nominative information (through a characterization of the corresponding persons) should be anonymized;

2. a patient appears in a certain database (e.g., for a medico-commercial study), only if it is obligatory or if he gives his consent;
3. in the same way, it is necessary to have the patient's consent when reversing this anonymity.
4. the anonymization procedure is based on a secret, the *patient anonymous identifier,* which is used to link uniquely the patient's anonymized medical data to the patient, while preserving his privacy.

Considering the identified needs, the use of a smart card is the most suitable means, and this is the most novel contribution of this paper.

At this step, we can easily suggest how the use of smartcards can help to meet these needs:

1. the smartcards are considered as sufficiently tamper-resistant to prevent, for instance, forging fake patient cards or cloning existing patient cards (in any case, the cost to forge or clone cards would be much higher than the expected profit);
2. with current smartcard technology, it is possible to anonymize nominative data;
3. by providing his smartcard, the patient gives his consent (at least implicitly); in other words, the use of a smartcard is suitable to guarantee the patient consent;
4. to keep secret the patient anonymous identifier, we suggest to generate it randomly, and to store it in the card; if in addition, we carry out the cryptographic transformations (anonymization and disanonymization) in the card, we guarantee that this secret (the identifier) is never transmitted out of the card. Section 4 gives details of our proposition.

In the next subsection we will first present what we mean by anonymization objectives and requirements. We will then refine our requirement analysis by giving details of some typical scenarios. Subsequently, for each scenario, we will identify some anonymization objectives and requirements. As suggested by our approach, our analysis will result in proposing a new anonymization procedure.

2.1 Anonymization objectives

An anonymization objective can be defined according to one of the three following properties, applied to the anonymization function [8]:

Reversibility: this corresponds to hiding data by encryption. In this case, from encrypted data, it is always possible to calculate the corresponding original nominative data (decryption), and conversely (encryption), by using the corresponding keys and algorithms.

Irreversibility: this is the property of anonymization. A typical example is to use a one-way hash function. Once replaced by anonymous codes, the original nominative data is no longer recoverable. However, in some cases, attacks by inference (or by dictionary) are able to re-identify the person if the anonymization is imperfect.

Inversibility: this is the case where it is, in practice, impossible to re-identify the person, except by applying an exceptional procedure. This procedure must be controlled by a highly trustworthy authority such as the medical examiner, the inspector-doctor or a trustworthy advisory committee. This authority can be seen as the privacy guarantor. Actually, it is a matter of a pseudonymization rather than anonymization, according to the common criteria terminology [9].

2.2 Anonymization requirements

The analysis of the requirements must take into account the system environment (users categories, attacks types, etc.). For instance, even if the information is anonymous, a malicious user can infer confidential information by using illegitimate deduction from external data. In this respect, two kinds of requirements must be imposed to any anonymization system: the *"linking"* requirements and the *"robustness"* requirements [8].

Linking allows associating (in time and in space) different anonymous data to the same person, without necessarily disclosing the real identity of that person. As mentioned in Figure 1, linking can be temporal (always, sometimes, never) or geographic (international, national, local, etc.).

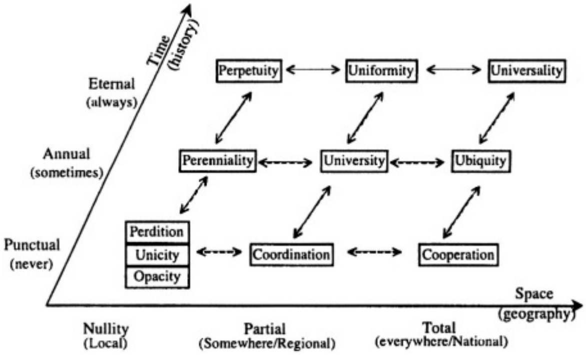

Figure 1. Network of the anonymization cases

The robustness requirements, concerning exclusively illicit disanonymization, can be divided into two distinct cases: robustness to reversion and to inference.

The *reversion robustness* concerns the possibility to inverse the anonymization function, for example if the used cryptographic techniques are not strong enough.

The *inference robustness* concerns data disanonymization by means of unauthorized recovery of nominative data. Generally, an inference can be:

deductive: it consists in inferring, mainly by first-order logic calculation, unauthorized information on the only basis of publicly available data;

inductive: if the reasoning that uses data explicitly stored in the information system is not sufficient to infer information, this reasoning can be completed by making some hypothesis on certain information;

probabilistic: it consists in inferring, by stating a set of various plausible assumptions, a secret information from valid available data.

This list is not exhaustive, and naturally, we can imagine other types of inference channels based on other types of reasoning.

2.3 Solution characterization

For a given scenario, once the privacy needs, objectives and requirement are defined, we have to characterize the most suitable solutions. In particular, we have to identify:

the *type of the solution* to develop: is it an organizational procedure, a cryptographic algorithm, a one-way function, or a combination of subsets of these solutions?

the *plurality of the solution:* do we need simple, double or multiple anonymization? Rationally, the choice is related to the type of threats considered against the anonymization function;

the *interoperability of the* solutions that are to be combined: *transcoding* (manually, for some continuity reasons) or *translating* (mathematically, for some consistency reasons) an anonymization system of anonymous identifiers into another anonymization system of anonymous identifiers; or *transposing* (automatically) several anonymization systems of anonymous identifiers into a unique anonymization system, in order to authorize or forbid the matching of anonymized data.

3. APPLICATION OF OUR METHODOLOGY TO TYPICAL HEALTHCARE EXAMPLES

3.1 Medical data transmission

The sensitivity of the information exchanged between healthcare providers (e.g., between the biology laboratories and the physicians) emphasizes the needs of confidentiality and integrity on transmitted data. Moreover, we need that only the legitimate addressee can receive and read the transmitted data. The use of an asymmetric (or hybrid) cryptographic system seems suitable [17].

The technique used should be reversible when duly authorized *(objective)* and robust to illegitimate reversion *(requirement).*

3.2 Professional unions

In France, for evaluation purpose, the physicians have to send to the professional unions data related to their activity. At first sight, a requirement is to hide patient's and physician's identities. However, when the purpose of use is to evaluate the physician's behavior (to assess care quality), it should be possible to re-identify the concerned physician. Our study of the French law [10] allowed us to identify the following anonymization objectives:

Inversible anonymization *(pseudonymization)* of the physician's identities: only an official body duly authorized to evaluate the physician's behavior can re-establish the real identities.

Inversible anonymization *(pseudonymization)* of the patient's identifiers: only welfare consulting doctors can reverse this anonymity.

In this way, the following risks are avoided:

attempts by a dishonest person to get more details than those necessary to his legitimate task in the system. For example, if the purpose is to study the global functioning of the system, it is not necessary to know the real identities (in accordance with the least privilege principle);

considering that the French law gives to patients the right to forbid the sharing of their information between several clinicians, the identified objectives aim to avoid privacy violation (inasmuch as patients could confide in some clinicians, and only in these clinicians).

3.3 PMSI framework

The *Information System Medicalization Program* (PMSI) aims at evaluating hospital information systems. Actually, it analyses the healthcare establishments activities in order to allocate resources while reducing budgetary inequality [11]. Given that the purpose is purely and simply medico-economic (and not epidemiologic), it is not necessary to know to whom a given medical information belongs *(anonymization).* On the other hand it is important to recognize that different data are related to the same, anonymous person even if they come from different sources at different times *(linkability).* Having said that, every patient must (always) have the same irreversible, anonymous identifier for the PMSI.

3.4 Statutory notification of disease data

Some diseases have to be monitored, through statutory notification, to evaluate the public healthcare policy (e.g., for AIDS) or to trigger an urgent local action (e.g., for cholera, rabies, etc.). Originally, patient records are nominative, but they are irreversibly anonymized before any transmission.

Various needs can be identified: prevention, care providing, epidemiological analysis, etc. The main objective is *anonymization* and *linkability.* Furthermore, universal linking, robustness to inversion, and robustness to inference are the main requirements.

In this respect, the choice in terms of protection must depend on these objectives. In fact, would we like to obtain an exhaustive registry of HIV positive persons? In this case, the purpose would be to know the epidemic evolution, and to globally evaluate the impact of prevention actions. Inversely, would we like to institute a fine epidemiological surveillance of the HIV evolution, from the infection discovery to the possible manifestation of the disease? In this case, the objective is to finely evaluate the impact of therapeutic actions, as well as a follow-up of certain significant cases.

This choice of objectives has important consequences on the nature of data to be collected, on the links with other monitoring systems, and consequently, on the access control policy and mechanisms.

Currently, we identify the following findings related to data impoverishment, to reduce inference risks:

Instead of collecting the zip code, it is more judicious to collect a region code. Obviously, a zip code could allow a precise geographic localization, resulting in identifying a small group of persons.

Instead of collecting the profession, we think that a simple mention of the socio-professional category is sufficient.

Instead of mentioning the country of origin it is sufficient to know if the HIV positive person has originated from a country where the heterosexual transmission is predominant.

3.5 Processing of medical statistical data

Nominative medical data should never be accessible, except when expressly needed for a course of treatment. This applies, in particular, to purely statistical processing and scientific publications. In this respect, not only such data should be anonymized, but also it should be impossible to re-identify the concerned person. Therefore, anonymization inversibility and robustness to inference are essential. Of course, everybody knows that, even after anonymization, identities could be deducted by a malicious statistician if he can combine several well-selected queries, and possibly, by complementing the reasoning by external information.

The problem of statistical database inference has been largely explored in previous works [12, 13]. In the reference [14, ch3] we list some example and solutions, but we believe that it is difficult to decide which solution is the most satisfying. In some cases, the solution could be to exchange the attributes values (in a certain database) so that the global precision of the statistics is preserved, while the precise results are distorted. The inherent difficulty in this solution is the choice of values to be permuted. Another solution could modify the results (of statistical requests) by adding random noise. The aim is to make request cross-checking more difficult.

3.6 Focused epidemiological studies

As mentioned earlier in the introduction, it is sometimes desirable to re-identify patients in order to improve care quality, especially in some focused studies such as cancer research protocols, genetic disease follow-up, etc.

To make this clearer, let us consider a simple example. Suppose that the patients of the category *C,* having undergone a certain treatment *Tbefore;* and that a certain study concludes that if these patients does not take the treatment *Tafter,* they will have a considerably reduced life expectancy. In such situations, it is necessary to re-identify the patients so that they take advantage of these new results. Having said that, we can conclude that this is a matter of an inversible anonymization. Of course, only authorized persons should be able to reverse the anonymity (e.g., consulting physician, medical examiner), and only when it is necessary.

In the case of cancer research protocols, the process starts by identifying the disease stage, then the protocol corresponding to the patient is identified, and finally, according to this protocol, the patient is registered in a regional, national or international registry. The epidemiological or statistical studies of theses registries could bring out new results (concerning patients following a certain protocol). In order to refine these studies and improve the scientific research, it is sometimes useful to re-identify the patients, to link some data already collected separately, and finally complement the results.

4. OUR SOLUTION

4.1 General scheme

Previously, we recommended that every anonymization needs a judicious prior analysis. This study must clearly and explicitly identify the security needs, objectives and requirements. After that, we have identified some scenarios and we have applied our approach to these scenarios. Now, we give shape to our analysis by developing a new solution that uses smartcards to better meet privacy needs and to satisfy the identified requirements.

First, in order to decide which view (specific form of data) is accessible by which user, our solution takes into consideration some parameters: the user's role, his establishment, and the purpose of use. Our main aims are to respect the least privilege principle as well as to make use of the legislation related to privacy, especially the European norms [15].

So as to do, before distributing anonymous data to end users, our solution suggests to cascade cryptographic processing (one after the other), carried out in different organizations. Of course, in each step, the transformation to carry out depends on the purpose of the use that follows. The outlines of our solution are first represented in Figure 2, then detailed and discussed in the following sections.

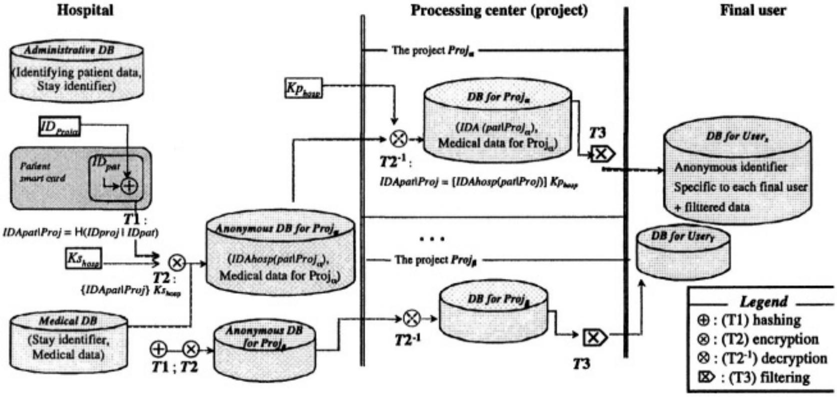
Hospital
Processing center (project)
Final user
Administrative DB
(Identifying patient data, Stay identifier)
$ID_{Proj\alpha}$
Patient smart card
ID_{pat}
T1 :
IDApat|Proj = H(IDproj | IDpat)
Ks_{hosp}
T2 :
{IDApat|Proj} Ks_{hosp}
Medical DB
(Stay identifier, Medical data)
T1 ; T2
Anonymous DB for $Proj_\alpha$
($IDAhosp(pat|Proj_\alpha)$, Medical data for $Proj_\alpha$)
Anonymous DB for $Proj_\beta$
Kp_{hosp}
The project $Proj_\alpha$
DB for $Proj_\alpha$
($IDA (pat|Proj_\alpha)$, Medical data for $Proj_\alpha$)
$T2^{-1}$:
IDApat|Proj = [IDAhosp(pat|Proj)] Kp_{hosp}
T3
...
The project $Proj_\beta$
DB for $Proj_\beta$
$T2^{-1}$
T3
DB for $User_x$
Anonymous identifier Specific to each final user + filtered data
DB for $User_y$
Legend
⊕ : (T1) hashing
⊗ : (T2) encryption
⊗ : ($T2^{-1}$) decryption
⌦ : (T3) filtering

Figure 2. The suggested anonymization procedure.

4.1.1 Transformations processed in healthcare establishments

In healthcare establishments (hospitals, clinics, etc.), three kinds of databases can be distinguished:

an administrative database, accessible to administrative staff (e.g., secretaries, reception staff), according to their role;

a medical database, accessible to clinical staff in charge of the patients;

several anonymous databases, each one containing the necessary and sufficient information for a certain project. A project is an entity that makes statistical or medico-economical data processing such as the PMSI, healthcare insurance companies, associations of diabetic persons, offices for medical statistics, research centers, etc.

The system must possess a well-defined access control policy, and this policy must be implemented by suitable security mechanisms. In the reference [16], we present the *Or-BAC (Organization-Based Access Control)* security model. Or-BAC offers the possibility to handle several security policies associated with different organizations. It is not restricted to permissions, but it also includes the possibility to define prohibitions, obligations and recommendations. In this respect, Or-BAC is able to specify policies developed for a large range of complex and distributed applications.

In our proposal, the transition from a medical database to an anonymized one (dedicated to a certain project) needs the application of two transformations, ***T*1** and ***T*2**.

*T*1*:* consists in calculating "$IDA_{pat|Proj}$", an anonymous identifier per person and per project. $IDA_{pat|Proj}$ is computed from the two identifiers "ID_{proj}" and "ID_{pat}", and characterizes the pair (patient, project):

ID_{proj} is the project identifier; it is known by the healthcare establishments that collaborate with this project;

ID_{pat} is the individual anonymous identifier of the patient; we state that this identifier should be kept under the patient control, on his personal *medical data smart card;* ID_{pat} is a random number generated uniquely for this patient, and is totally independent from the social security number; a length of 128 bits is sufficient to avoid collisions (the risk that two different persons have the same identifier).

In the healthcare establishment, at the time of supplying data to the anonymous databases (per project), the user (i.e., the hospital employee) transmits ID_{proj} (the project identifier) to the card. The card already contains ID_{pat} (the patient anonymous identifier). By supplying his card, the patient gives his consent to transmit his medical data as part of this project. The *T1* procedure, run by the smart card, consists in applying a one-way hash function (i.e., SHA) to the concatenated set (ID_{proj} | ID_{pat}):

$$\textbf{(T1)} \qquad IDA_{pat|Proj} = \mathrm{H}(ID_{proj} \mid ID_{pat})$$

By generating the fingerprint $\mathrm{H}(ID_{proj} \mid ID_{pat})$, *T*1 aims at the following objectives:

the patient data appear in a certain database only if it is obligatory or if he has given his consent by producing his patient data card;

$IDA_{pat|Proj}$ does not use any secret whose disclosure would undermine the privacy of other persons (as opposed to the use of a secret key that would be used for all the patients). In addition, since $IDA_{pat|Proj}$ calculation is run into the card, ID_{pat} remains into the card; it is never transmitted outside, and it is only used for creating an anonymous database entry (in the hospitals);

since ID_{proj} is specific to each project, the risks of illicit linkage of data belonging to two different projects are very low; moreover, the anonymous databases (per project) are isolated from external users, and so, can be protected by strict measures of access control;

knowing that $IDA_{pat|Proj}$ is always the same for the pair (patient, project), every project can *link* data concerning the same patient, even if they are issued by different establishments or at different times, as long as they concern the project.

Nevertheless, the transformation *T*1 does not protect against attacks where attackers try to link data held by two different hospitals. To make this clearer, let us take an example where a certain patient Paul has been treated in the hospitals $Hosp_A$ and $Hosp_B$. In each of these two hospitals, Paul has consented to give his data to the project $Proj_\alpha$. Let us assume that Bob, an $Hosp_B$ employee, knows that the fingerprint **X** $(=IDA_{Paul|Proja})$ corresponds to Paul, and that Bob obtains (illicitly) access to the anonymous database held by $Hosp_A$ and concerning $Proj_\alpha$. In this case, the malicious user Bob can easily establish the link between Paul and his medical data (concerning $Proj_\alpha$) held by $Hosp_A$ and $Hosp_B$.

In order to face this type of attacks, a cryptographic asymmetric transformation (**T2**) is added. Thus, before setting up the anonymous databases (specific to each project), the hospital encrypts (using an asymmetric cipher) the fingerprint $IDA_{pat|Proj}$ with the key Ks_{hosp} specific to the hospital; (the notation "{M}K" indicates that M is encrypted with key K):

$$\textbf{(T2)} \qquad IDAhosp(pat|Proj) = \{IDA_{pat|Proj}\}\ Ks_{hosp}$$

If we take again the previous scenario, the malicious user Bob cannot re-identify the patients because he does not know the decryption key Kp_A. In fact, each hospital holds its key Ks_{hosp}, while Kp_{hosp} is held only by the projects.

It is easy to observe that the two transformations (*T*1 and *T*2) allow having a great robustness against attacks attempting to reverse the anonymity (in particular by linking) in an illicit manner.

The procedure, carried out in the hospitals, remains very flexible. Indeed, if two hospitals ($Hosp_a$ and $Hosp_b$) decide to merge someday, it is easy to link data concerning every patient that has been treated in these hospitals. In fact, each hospital decrypts its data with its key Kp_{hosp}, and then encrypts the result by $Ks_{hosp_{ab}}$ the new hospital private key. If $IDAhosp_a(pat|Proj)$ (respectively $IDAhosp_b(pat|Proj)$) designates an anonymous identifier in $hosp_a$ (respectively $hosp_b$), and "[]K" designates a decryption with K:

The processing carried out on the former data of $hosp_a$ is:

$$\{\ [IDAhosp_a(pat|Proj)]\ Kp_{hospa}\ \}\ Ks_{hospab}\ ;$$

The processing carried out on the former data of $hosp_b$ is:

$$\{\ [IDAhosp_b(pat|Proj)]\ Kp_{hospb}\ \}\ Ks_{hospab};$$

In this way, the resulting fingerprints are the same in the two hospitals (for each anonymous database associated to a certain project).

4.1.2 Transformations carried out when projects receive data

Data contained in the anonymous databases (in the hospitals) undergoes transformations that depend on $IDA_{proj|pat}$ and on Ks_{hosp}. Every processing center (project) decrypts received data by using Kp_{hosp}:

$$[IDAhosp(pat|Proj)]\ Kp_{hosp}$$

according to (T2), $= [\ \{IDApat_{|Proj}\}\ Ks_{hosp}\ ]\ Kp_{hosp} = IDApat_{|Proj}$

The processing center thus retrieves, for each patient, the persistent anonymous identifier dedicated to the project, $IDApat_{|Proj}$, i.e., an information that is sufficient and necessary to link data corresponding to each patient, even if they come from different establishments at different times.

4.1.3 Transformations carried out before the distribution to the end users

Before their distribution to the end users (scientist researchers, web publishing, press, etc.) the information can undergo a targeted filtering. For instance, this can be done by applying data aggregation or data impoverishment.

If an additional security objective is to control if end users can link information, it is advisable to apply another anonymization (e.g., by MD5) with a secret key $Kutil_{|proj}$ generated randomly.

$$IDApat_{|util} = \mathrm{H}(IDApat_{|Proj}\ |\ Kutil_{|proj})$$

In accordance to needs, this transformation corresponds to two different cases:

If the aim is to allow the full time linking (per project for that particular user), the key $Kutil_{|proj}$ has to be stored by the processing center, so that it can reuse it each time it transmits information to this end user.

Inversely, if the center wishes to forbid users to link data transmitted at different times, the key $Kutil_{|proj}$ is randomly generated just before each distribution.

4.2 Discussion

By using the personal smartcard, the solution that we suggest guarantees the following benefits:

The smartcards are sufficiently tamper-resistant.

The secret as well as the anonymization and the disanonymization procedures are held under the patient control. Not only the patient identifier is never transmitted out of the card, but also the generation of project-specific anonymous identifiers is carried out within the card.

The patient's consent must be provided for each non-obligatory, but desirable, utilization of his anonymized data. Indeed, the patient medical data can appear in a certain database only if, by supplying his card, the patient gives his consent to exploit his medical data as a part of a certain project (e.g., epidemiological research concerning a rare disease). The same goes when reversing the anonymity: the patient consent is necessary. Let us take the example where the end user (e.g., researcher in rare diseases) discovers important information that necessitates re-identifying the patients. At first, it sends back results to the hospitals participating to the concerned project (e.g., a given orphan disease study). Only if the patient produces his medical data card (which implies that he gives explicitly his consent), it is possible to calculate $IDA_{pat|Proj} = \mathrm{H}(ID_{proj} \mid ID_{pat})$ and $IDAhosp(_{pat|Proj}) = \{IDA_{pat|Proj}\}_{Kshosp}$, and so, to establish the link between the patient, his anonymous identifiers, and his medical data. A simple (and automatic) comparison between the anonymous identifier and the inversion list, would allow triggering an alarm. This alarm asks the patient if he wants to consult the results.

The solution resists dictionary attacks that could be led in different organizations: hospitals, processing centers and end users.

The combination of the suggested anonymization procedure with access control mechanisms satisfies the non-inversibility requirement as well as the least privilege principle.

It is possible to merge data belonging to several establishments without compromising security nor flexibility;

In accordance with European legislation, our solution takes the purpose of use into account. Moreover its fine-grain analysis allows to easily adapt it to different sector needs (e.g., social domain, demographic projects, other scientific research areas, etc.).

Currently, French hospitals anonymize patient identities by using a one-way hash coding based on the standard hash algorithm (SHA). The principle is to ensure an irreversible transformation of a set of variables that identify an individual (last name, first name, date of birth, sex). In order to link all the information concerning the same patient, the anonymous code obtained is always the same for the given individual (and for all uses). Two keys have been added before applying SHA. The first pad, k1, is used by all senders of information as follow "$Code_1 = \mathrm{H}(k_1 \mid Identity)$"; and the second, k2, is applied by the recipient "$Code_2 = \mathrm{H}(k_2 + Code_1)$". Nominal information is therefore hashed twice, consecutively with these two keys. The aim of pad k1 (resp. k2) is to prevent attacks by a recipient (resp. a sender) (Figure 3).

However, this protocol is both complex and risky: the secret key should be the same for all information issuers (clinicians, hospitals) and stay the same over time. Moreover, this key must always remain secret: if this key is corrupted, the security level is considerably reduced. It is very difficult to keep secret during a long time a key that is largely distributed. This means that new keys have to be distributed periodically. The same applies when the hash algorithm (or the key length) is proven not sufficiently robust any more. But, how can we link all the information concerning the same patient before and after changing the algorithm or the key? If this problem occurs, the only possible solution consists in applying another cryptographic transformation to the entire database, which may be very costly.

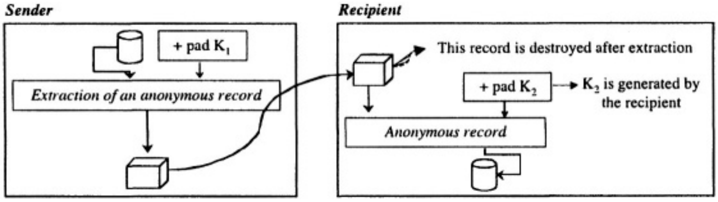
Sender
Recipient
+ pad K_1
Extraction of an anonymous record
This record is destroyed after extraction
+ pad K_2
K_2 is generated by the recipient
Anonymous record

Figure 3. Outline of the current French hashing procedure

Inversely, in our solution, the identifiers (ID_{proj}, ID_{pat}, $IDA_{pat|Proj}$ and $IDA_{pat|util}$) used in the various transformations are located in different places. Similarly, the keys (Ks_{hosp}, Kp_{hosp}) are held by different persons. Indeed, ID_{proj} concerns a unique project; the pair (Ks_{hosp}, Kp_{hosp}) is specific to one hospital; $IDA_{pat|util}$ is dedicated to a single end user. It is therefore practically impossible to make illicit disanonymization. Moreover, ID_{pat} is specific to one patient, and only held on his card. So even if a certain ID_{pat} (corresponding to Paul, for example) is disclosed (which is not easy!), only Paul's privacy could be endangered (but not all the patients' privacy, as it is the case in the current French procedure) and only for certain projects.

Furthermore, according to the security needs of the studied cases, we suggest to complement our solution by other technical and organizational security mechanisms:

the access to data has to be strictly controlled; a well-defined security policy must be implemented by appropriate security mechanisms (hardware and/or software);

sometimes, it is advisable to use thematic anonymizations, so that even if a certain user succeeds in breaking the anonymity, risks (particularly inference risks) are limited;

in some particular contexts, it is more efficient to completely remove identifying data from medical data.

as complementary deterrent or repressive measures, it is recommended to control the purpose of use by implementing intrusion detection mechanisms; in particular, these mechanisms should easily detect sequences of malicious requests (illicit inferences, abuse of power, etc.).

CONCLUSION

In an electronic dimension that becomes henceforth omnipresent, this paper responds to one of the major recent concerns, fathered by the new information and communication technologies: the respect of privacy.

In this framework, we firstly analyzed the anonymization in the medical area, by identifying and studying some representative scenarios. Secondly, we have presented an analytic approach putting in correspondence anonymization functionalities and adequate solutions. Finally, we suggested a new procedure adapted to privacy needs, objectives and requirements of healthcare information and communication systems. This fine-grain procedure is generic, flexible and could be adapted to different sectors. The use of the smartcards in this procedure responds to some security needs.

Although this solution suggests successive anonymizations, the cryptographic mechanisms that it uses are not expensive in terms of time and computation resources, and are compatible with current smartcard technology. We are currently implementing a prototype of this solution, using Java Cards, and we will soon be able to measure the performance and complexity of a real application.

REFERENCES

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

[2] 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.

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

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

[5] Loi 78-17 du 6 Janvier 1978 relative a 1'Informatique, aux fichiers et aux libertiés, Journal officiel de la République française, pp. 227-231, décret d'application 78-774 du 17 juillet 1978, pp. 2906-2907.

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

[7] A. Pfitzmann, M. Köhntopp, "Anonimity, Unobservability, and Pseudonymity – A Proposal for Terminology", *International Workshop on Design Issues in Anonymity and Unobservability,* Berkley, CA, USA, July 25-26, 2000, Springer.

[8] Trouessin, G (1999). "Dependanility Requirements and Security Architectures for Healthcare/Medical Sector", 18th *International Conference SAFECOMP'99,* Toulouse, France, September 1999, Springer, pp. 445-458.

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

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

[11] 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 1'emploi et de la solidarité France.

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

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

[14] A. Abou El Kalam, *"Modèles et politiques de sécurité pour les domaines de la santé et des affaires sociales",* Thèse de doctoral, Institut National Polytechnique de Toulouse, 190 pp., 4 December 2004.

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

[16] A. Abou El Kalam, P. Balbiani, S. Benferhat, F. Cuppens, Y. Deswarte, R. El-Baida, A. Miège, C. Saurel, G. Trouessin "Organization-Based Access Control", *4th International Workshop on Policies for Distributed Systems and Networks (Policy'03),* Como, Italy, 4-6 June 2003, IEEE Computer Society Press, pp. 120-131.

[17] A. *Menezes,* P. C. Van Oorshot, S. A. Vanstone, *"Handbook ofApplied Cryptography",* 1997, CRC press, ISBN : 0849385237, pp. 780.