International Journal of Computer Networks and Communications Security

VOL. 3, NO. 5, MAY 2015, 208–219
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

SECURITY & PRIVACY BY DESIGN: A new approach for Healthcare Information and Communication Systems

Anas ABOU EL KALAM[1], Jean-Philippe LEROY[2], Larbi BESSA[3] and Jean-Marie MAHE[4]

[1, 2, 3, 4] IPI –LISER / Propedia, Paris, France

E-mail: [1]aabouelkalam, [2]jpleroy, [3]l.bessa, [4]jmmahe}@groupe-igs.fr

ABSTRACT

Nowadays, more and more applications use sensitive and personal information. Subsequently, respecting citizens' privacy while preserving information security is becoming extremely important. Initially, deploying security mechanisms as well as Privacy-Enhancing Technologies (PETs) was seen as the solution. Today, we realize that a more substantial approach is required, taking into account the security and privacy needs from the earlier steps of the system specification. Dedicated to this issue, this paper is organized as follows: after defining the topic through several examples, this paper analyzes the most typical anonymization procedures used in various countries and presents the main privacy-related concepts. Then, it suggests a rigorous approach to define suitable anonymization solutions and mechanisms through the needs, objectives and requirements. Afterwards, a representative range of scenarios is presented and confronted to the approach already described. Finally, a new generic procedure to anonymize and link identities is suggested. Details about the implementation and analysis of our solution are also presented. Our approach takes the purpose of use into consideration, guarantees the citizen's consent, resists dictionary attacks, respects the least privilege principle and thus fulfills the European legislation requirements. Even if our approach is applied in this paper to healthcare examples, it could also be suitable to every system with security and privacy needs.

Keywords: *Anonymization, Security, Privacy, Health Care, Electronic Medical Records.*

1 INTRODUCTION

For the time being, we can assert that international [1], American [2,3,4] and European legislations are not only worried about protecting personal and nominative data, but also aim at forbidding files linkage [5, 6, 7, 8]. Moreover, in many organizations, privacy is considered as a purely legal issue; and a big gap persists between its identification and implementation. Worst, security and privacy are sometimes considered as separate issues, and the deployed security mechanisms often threaten privacy. For example, in healthcare systems, authentication and traceability mechanisms are used to identify reliably the patients; on the other hand, strong security may endanger the patient's privacy.

To satisfy the privacy-related legislations, countries and institutions uses classical Privacy-Enhancing Technologies (PETs) such as anonymization [9, 10, 11, 12].

However, classical mechanisms are not satisfying in complex systems as it is sometimes possible to identify a person by linking non-nominative data, by breaking the privacy mechanisms or by using inference techniques. For instance, the age, the sex and the month of discharge from hospital are enough to identify the patient in a limited population. Likewise, it is commonly known that two childbirth dates is enough to identify a woman in a sizeable population.

In this paper, we explain that the privacy (as well as security) should be studied from the earlier phases of the system specification, taking into account the needs, the objectives and the requirements. We thus propose a systematic methodology that progressively derives the privacy

related mechanisms, and we apply it to the healthcare system.

Subsequently, this paper is organized as follows: Section 2 explains classical solutions and shows their main drawbacks. To overcome these limitations, Section 3 proposes a systematic methodology that first analyzes the privacy needs, specifies the privacy objectives and finally derives the privacy requirements. Once these steps achieved, it would be possible to identify the suitable mechanisms that satisfies the needs and overcomes the risks. To show the usability of our methodology, we apply it to healthcare information and communication systems. Subsequently, we derive in Section 4 a generic solution based on the main steps of our methodology. Afterward, a security analysis of our work is proposed in Section 5. Finally, Section 6 concludes our work and perspectives.

2 CLASSICAL SOLUTIONS

Healthcare organizations represent excellent examples of systems with strict security and privacy needs. In fact, in order to make the accurate diagnoses and provide the best treatment, patients naturally provide and share sensitive personal information with their healthcare professionals. This information may also be shared with others, such as insurance companies, pharmacies, researchers, and employers, for many reasons. If patients are not confident that this information will be kept confidential, they will not be forthright and reveal accurate and complete information. Moreover, if healthcare providers are not confident that the organization that is responsible for the healthcare record will keep it confidential they will limit what patients add to the record. Either of these actions is likely to result in inferior healthcare. Subsequently, several laws and rules have been published to protect the privacy and security of personal health information. To enforce these legislations, each country has taken the necessary measures and deployed the suitable measures.

For instance, several French hospitals use an anonymization protocol [13] that transforms patient identities by using a one-way hash function (SHA). The principle is to ensure an irreversible transformation of a set of identifying variables (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.

However, this procedure is vulnerable to dictionary attacks (e.g., by comparing hashed known identities with the code assigned to a particular patient). In order to avoid such attacks, two keys have been added before applying SHA. The first pad, k1, is used by all senders of information as follow “Code1 = H(k1 | Identity)”; and the second, k2, is applied by the recipient “Code2 = H(k2 + Code1)”. 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).

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 it is corrupted, the security level is considerably reduced. It’s very difficult to keep a key secret during a long time, especially if it is largely distributed. This means that new keys have to be distributed periodically. The same problem occurs 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 when it becomes necessary to change 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.

In Germany, the National Cancer Registry (GNCR) is used for collecting medical statistics related to cancer. The procedure of the population-based cancer registration is carried out in two steps by two institutions [14]. In the first step, the Trusted Site collects the tumor-related data recorded by doctors, dentists or Follow-up Organization Centers. The Trusted Site anonymizes the patient’s personal data by an asymmetric procedure, e.g., a hybrid IDEA-RSA encoding: the identifying data is encrypted with an IDEA session key, generated randomly; the IDEA key is then ciphered by a public RSA key. To allow an unambiguous assignment of additional information to the correct patient record, a control number (a special kind of pseudonym) is generated, using different attributes of the patient’s personal data. This control number is generated by using a one-way hash function (MD5) and a symmetrical cipher (IDEA). To allow the assignment of data from the different federal Lander, the control number procedure and key are unique all over Germany ("Linkage Format"). The Trusted Site transfers both the encrypted patient-identifying data and the epidemiological plaintext data to the Registry Site. The latter stores the record in the register database and brings together different records belonging to one patient. After matching the data, a random number is added to the control number and the

result is symmetrically encrypted by IDEA ("Storage Format"). To match new records, the control numbers must be transformed back from the "Storage Format" to the "Linkage Format".

In Switzerland, the Federal Statistics Office (SFSO) is responsible for collecting medical data. Information on the diagnoses and on the corresponding treatments is recorded for all patients. To preserve the patient's privacy, the SFSO has contacted the Swiss Federal Section of Cryptography (SFSC) to find a cryptographic expert capable of finding a solution to this problem. This section gives the outline of the SFSC's analysis [15]. First, it is not necessary to know to whom a given medical record belongs; however the SFSO needs to recognize whether two different records actually belong to the same patient. This is crucial in order to follow the history of the patients.

Basically, data can be split into two categories: medical and non-medical data. The SFSC suggests to replace identifying data (date of birth, sex, last name and first name) by a hash-computed personal code (fingerprint), called "anonymous linking code": fingerprint = Hash[ID-Var].

When information has to be transmitted by an hospital, a session key "c" is generated by hospital computers; c is then used to encrypt the fingerprint during the transmission to the SFSO: IDEA [fingerprint]c; a public key cryptosystem (RSA) is used to transmit the session key RSA[c]E.

After reception, c is calculated by using the SFSO private key D; the encrypted fingerprints are decrypted, and uniformly re-encrypted by the fragmented secret key K of the SFSO: they become the "anonymous linking codes", used as pseudonyms.

The cryptographic transformations carried out in the SFSO should be done in a "Black-box" (intermediate steps of these transformations should never be visible to the SFSO operators). However, how can we be sure that the secret key, the private key and the fingerprints are never recorded in a storage medium? We think that these steps (computation phases) should be done in a well-protected hardware module. Tamperproof access control mechanisms, possibly hardware, could improve the protection. The aim is that only trustworthy persons, acting together, can carry out the composite operation of calculus. The authorization server gives them the corresponding rights without disclosing the fingerprint nor the secret keys k and c.

Unfortunately, even if the European legislation attaches importance to the patient's consent (before using his personal data), the solutions presented above do not explain the technical procedures carried out to make this consent mandatory. In the same way, even if in some cases (e.g., to improve care quality) it could be important to re-identify patients (i.e., disanonymization), these solutions do not detail if it is possible (and if yes, how?). Furthermore, in all these anonymization schemes, only the patient's identity is anonymized, the medical data are in plaintext, and thus vulnerable to attacks by inference. Table 1 summarizes the techniques presented above and lists some of their weaknesses.

Table 1: summary of some existing anonymization procedures.

Solution	Description	Main weaknesses
France	• One-way hash function • Secret key	• The difficulty to keep a key secret during a long time, especially because it is largely distributed • The patient's consent? • Disanonymization, if necessary?
Germany	• Session key *c* • $IDEA[ID\text{-}data]_c$ • $RSA[c]_E$	• The patient's consent? • Disanonymization?
Suitzerland	• Session key *c* • Fingerprint = $Hash[ID\text{-}data]$. • $IDEA[fingerprint]_c$; • $RSA[c]_E$	• How can we be sure that the private key and the fingerprints are never recorded in a storage medium? • The patient's consent? • Disanonymization?

3 ANALYTIC APPROACH

The previous section shows that most existing solutions are developed empirically, and thus have some weaknesses. In particular, we believe that before calling for technical or organizational solutions, we should first develop an analytic approach based on needs identification and risks analysis. In fact, privacy analysis can be expressed with respect to two levels of abstraction:

- The request: a set of needs expressed by the legislation, hospitals, citizens, etc. Basically, we specify in this step "what" is expected from the "client / citizen / user / law / project owner" point of view?

- The response in the form of security functionalities, mechanisms and solutions to implement. Basically, this step expresses the "implementer" point of view.

However, we believe that between the request and the answer, a rigorous privacy analysis should be specified, done and justified. In particular, it is necessary to clearly identify:

- Privacy objectives: security level to reach, information to protect, threats to avoid, etc.
- Security requirements: formalization of security needs, identification of functionalities, characterization of the solution, etc.

3.1 Privacy needs

The privacy needs represent the user expectations; they depend on the system, the environment, the legislation, etc. generally, their form is neither very explicit nor very simple to formalize. We suggest that the user needs should, at least, be expressed regarding the "Common Criteria" privacy-related functionalities [16]:

- Anonymity is the property of being "not identifiable" within a set of subjects;
- Pseudonymity adds accountability to anonymity; it ensures that a person may use a resource or a service without disclosing his identity, yet be accountable for that use;
- Unlinkability ensures that it is not possible to distinguish if two items are related or not;
- Unobservability ensures that someone may use a resource or service without other parties being able to observe whether the resource or service is being used.

Examples of privacy needs in health care systems could be:

- both directly nominative and indirectly nominative information (e.g., address) should be anonymized;
- a patient appears in a particular database (e.g., for a medico-commercial study) only if he gives his consent;
- Ensure unlinkability of files belonging to a particular patient except for the attending physician;
- Ensure the unlinkability of patient records and reducing the risks by limiting the collected personal data to the strict minimum;
- Reduce the severity of risks by ensuring that personal data will not be kept longer than necessary according to the national legislation;
- Allow citizens to oppose the use of their personal data;
- …

3.2 Anonymization objectives

In general, the security objectives are results of a privacy risk analysis; they that specify the security level to reach, the information to protect, the threats to avoid, the origin, nature and scenarios of threats, etc.

For instance, the objectives should specify:

- Who can access the data (and how it is approved by the data owner), what conditions and when that access is allowed?
- How data is collected and is authorized to process?
- Who fixed the period during which the data must be available?
- In which condition the desanonymization is possible, who trigger and who is responsible for the desanonimization (the judge, forensic scientist, coroner, medical examiner, …).
- Who fixed the evidence to provide, how accountability is ensured in case of pseudonymization
- etc.

Hence, the answer to this kind of questions should define the privacy objective according to one of the following properties, applied to the anonymization function:

- Reversibility: mainly, hiding data by encryption. In this case, from encrypted data, it is always possible to retrieve the corresponding original nominative data, and vice versa.
- Irreversibility: the property of anonymization. The typical example is a one-way hash function. Once replaced by anonymous codes, the original nominative data are no longer recoverable.
- Inversibility: this is the case where it is impossible to re-identify the person, except by applying an exceptional procedure restricted to duly authorized users. This exceptional procedure must be done under surveillance of a high trustworthy authority like 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 pseudonymisation in the common criteria terminology [16].

3.3 Anonymization requirements

At this stage, the analysis of the needs is first completed by information about the system environment (users categories, attacks types, etc.). To derive the security requirements, the needs are then represented with a non-ambiguous semantics (e.g., formal system). For instance, even if the information is anonymous, a malicious user can deduce confidential information by using illegitimate types of reasoning. In this respect, two kinds of privacy requirements must be imposed to any anonymization system: the "linking" requirements and the "robustness" requirements.

In particular, linking allows associating one or several pseudonyms to the same person. At this step we should thus clearly specify the duration and scope of the linkability and/or observalibily (of data, resources, processes, etc.) – already identified in the previous step. For example, should a particular data remain "always / sometimes / .. / never" anonymous (temporal characterization of the anonymization function)? And this anonymization should be maintained at the local / institutional / regional / national / … / international level (special and geographical characterization of the anonymization function)?

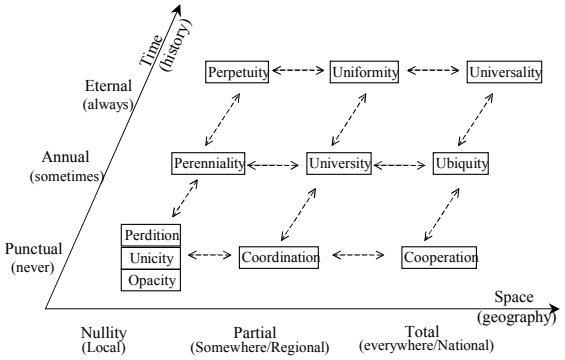

Fig. 1. Example of spatial & temporal characterization of the anonymization function.

Besides, the robustness requirements concerns illicit disanonymization. We distinguish robustness to reversion and to inference. The reversion robustness concerns the possibility to inverse the anonymization function. For instance, very critical data and procedures necessitate strong (and costly) cryptographic techniques with in-depth defense. The inference robustness concerns data disanonymization by means of unauthorized computation, e.g., by inference. Basically, we identify several kinds of inference:

- deductive: it consists in inferring, mainly by first-order logic calculation, unauthorized information on the only basis of publicly available data;
- Inductive: when the conventional reasoning that uses information 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, an unexpected secret information from valid available data.

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

Besides that, the anonymization requirements should specify:

- The type of 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 to implement: do we need simple, double or multi-anonymization?
- The interoperability of the solutions that are to be combined: transcoding (manually) or translating (mathematically) an anonymization system into another one; or transposing (automatically) several anonymization systems into a unique anonymization system.

4 APPLICATION OF OUR METHODOLOGY

4.1 Medical data transmission

The sensitivity of the information exchanged between healthcare providers emphasizes the needs of confidentiality and integrity on transmitted data. Moreover, we need only the legitimate addressee to receive and read the transmitted data. Therefore, the technique used should be reversible when duly authorized (objective) and robust to illegitimate reversion (requirement). Consequently, the use of an asymmetric (or hybrid) cryptographic system seems suitable [17].

4.2 Professional unions

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 and physician's identities. However, when the purpose of use is to evaluate the physician's behavior, it should be possible to re-identify the concerned physician. Our study of the European law 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).

4.3 ISMP Framework

The Information System Medicalization Program (ISMP) aims at evaluating hospital information systems in France; it analyses the healthcare establishments activities in order to allocate resources while reducing budgetary inequality. 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). So, every patient must (always) have the same irreversible anonymous identifier for the ISMP (the privacy objective). Furthermore, the robustness to inferences and to reversions (of the anonymization function) is essential (the privacy requirement).

4.4 Processing of Statutory Notification Disease Data

Some diseases have to be monitored, through statutory notification, to evaluate the public healthcare policy (e.g., AIDS) or to trigger an urgent local action (e.g., for cholera, rabies, etc.).

Various needs can be identified: prevention, care providing, epidemiological analysis, etc. The main objectives are 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. 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 second case, the objective is to finely evaluate the impact of therapeutic actions, as well as a follow-up of some 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, a simple mention of the socio-professional category could be sufficient.
- Instead of mentioning the country of origin it is probably sufficient to know if the HIV positive person has originated from a country where the heterosexual transmission is predominant.

4.5 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 should such data be anonymized, but it should also be impossible to re-identify the concerned person. Therefore, anonymization irreversibility 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 [13, 14]. In the reference [15, 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 particular 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 crosschecking more difficult.

4.6 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. Supposing that the patients of the category C, having undergone a treatment Tbefore; and that a particular study concludes that if these patients do 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. So this is a matter of an inversible anonymization. In fact, only authorized persons should be able to reverse the anonymity (e.g., consulting physician, medical examiner), and only when it is necessary.

In other cases, e.g. 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 added in a national or international registry. The epidemiological or statistical studies of theses registries could bring out new results (concerning patients following a particular protocol). In order to refine these studies and improve the scientific research, it is sometimes useful to re-identify the patients, so as to link some data already collected separately, and finally complement the results.

5 A GENERIC SOLUTION

Previously, we recommended that every anonymization necessitates a judicious prior analysis. This study must clearly and explicitly identify the security needs, objectives and requirements. After that, we have identified some representative scenarios and we have applied our approach to them. Now, we give shape to our analysis by developing a generic solution that can be adapted in order to satisfy most of the raised requirements.

First, in order to decide which data is accessible by which user, our solution takes into consideration some parameters such as the user's role, his establishment, and the purpose of use. Our major aims are to respect the least privilege principle as well as to make use of the legislation related to the privacy.

So as to do, before data distribution to the final users, our solution combines some organizational and technical procedures. In each step, the transformation to carry out depends on the use that follows. The outlines of our solution are first represented in Fig. 1, then detailed and discussed in the following subsections.

5.1 Transformations processed in healthcare organizations

Basically, 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);
- A medical database, accessible to clinical staff in charge of the patients; and
- Several anonymized databases, each one containing the necessary and sufficient information for a particular project. A project is an entity that makes statistical, epidemiological or medico-economical data processing such as the PMSI, the healthcare insurance, associations of diabetic persons, offices for medical statistics, research centers, etc.

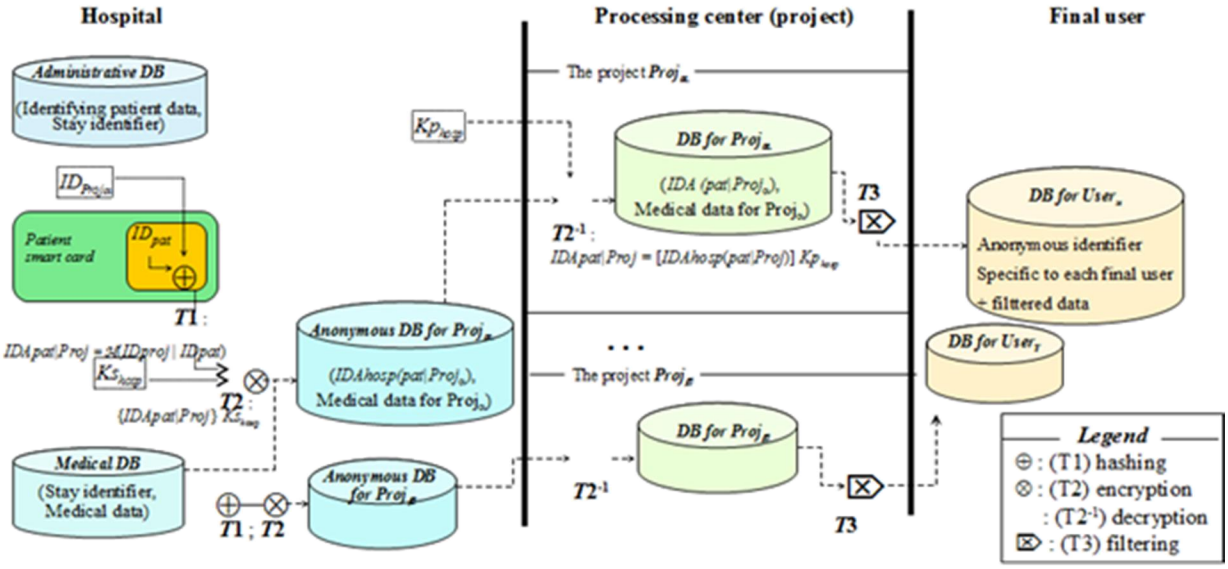

Fig. 2. The suggested anonymization procedure

It is worth noting that the transition from a medical database to an anonymized one (intended for a project) necessitates the application of two transformations (*T*1 and *T*2).

***T*1**: consists in calculating "*IDApat|Proj*", an anonymous identifier per person and per project. We believe that, contrary to some existing solutions (cf. Section 2), for a particular patient, it is not necessary to consider the same anonymous identifier for every use. That is why in our solution: (1) the possible thematic of uses are broken down into several projects[1]; (2) an identifier "*IDproj*" is associated to each project; (3) to characterize the pair (patient, project), *IDApat|Proj* is then computed from the two identifiers *IDproj* and "*IDpat*":

- "*IDproj*" is the project identifier; it is held by the healthcare establishments that collaborate with this project;
- "*IDpat*" is the individual anonymous identifier of the patient; we suggest that this identifier is held under the patient control, for example on his personal *medical data smart card*; *IDpat* is a random number generated uniquely for this patient, and is totally independent from the social security number; a length of 128 bits

[1] Obviously, this step (the identification of the projects) necessitates a prior needs analysis.

seems sufficient to avoid collisions[2] (the risk that two different persons have the same identifier).

In the healthcare establishment, when entering data into anonymous databases (per project), the user (i.e., the hospital employee) sends *IDproj* (the project identifier, concerned by this database) to the card. The card already contains *IDpat* (the patient identity). It is important to note that, by supplying his card, the patient gives his consent to exploit 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 (*IDproj* | *IDpat*):

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

By generating the fingerprint $\mathcal{H}(IDproj \mid IDpat)$, *T*1 aims at the following objectives:

- the patient appears in a database, only if it is compulsory or if he gives his consent by providing his patient data card (*IDpat* is present only in the card, and is never transmitted out of the card), for instance for a medico-commercial study;
- *IDApat*|*Proj* does not use any secret of which the divulgation undermines other person's privacy (contrary to the use of a secret key, common for all the patients, such as in the actual French solution, cf. Section 2.1.1). In addition, since *IDApat*|*Proj* calculation is run into the card, *IDpat* remains into the card; it is never transmitted outside, and it is only used for creating an anonymous database entry (in the hospital);
- Since *IDproj* 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 then, can be protected by strict measures of access control;
- Knowing that *IDApat*|*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 intrusions where attackers link data held by two different hospitals. To make this clearer, let us take an example where a 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 $IDAPaul|Proj\alpha$ 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) *IDApat*|*Proj* with the key Ks_{hosp} specific to the hospital; (the notation "{M}K" indicates that M is encrypted with key K):

(T2) $IDAhosp(pat|Proj) = \{IDApat|Proj\}\ Ks_{hosp}$

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[3]. 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) bring 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 $Kshosp_{ab}$ 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 (ancient) data of $Hosp_a$ is:

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

- The processing carried out on the former (ancient) data of $Hosp_b$ is:

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

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

[2] The proof is based on the birthday problem.

[3] K_{pA} is known by all the project centers that $Hosp_A$ cooperates with, but is not "public". On the other side, K_{SA}, the corresponding "private" key, is known only by $Hosp_A$.

5.2 Transformations carried out upstream from processing centers

Data contained in the anonymous databases (in the hospitals) undergoes transformations that depend on *IDAproj|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 finds the information that is sufficient and necessary to its processing. Since this information is associated to *IDApat|Proj*, each project can link data corresponding to the same patient.

5.3 Transformations carried out before the distribution to the final users

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

If, in addition, the security objective is to forbid final users to link information, it is advisable to apply another anonymization (e.g., by MD5) with a secret key *Kutil|proj* generated randomly.

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

In accordance to needs, this last case corresponds to two different processes:

- 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 this same key when transmitting information to the same final user;
- Inversely, if the center wishes to forbid users linking data, the key is randomly generated just before each distribution.

6 DISCUSSION

The suggested generic solution brings mainly the following benefits:

- Every step (technical or organization procedure) necessitates a judicious prior analysis of privacy risks, needs, objectives and requirements.
- The anonymous patient identifier differs from a project to another.
- The patient's consent must be provided for each non-compulsory, but desirable, utilization of his anonymized data.
- The identifiers (*IDproj*, *IDpat*, *IDApat|Proj* and *IDApat|util*) used in the various transformations are located in different places; similarly, the keys (*Kshosp*, *Kphosp*) are held by different persons. Indeed, *IDproj* concerns a unique project; *IDpat* is specific to one patient, and only held on his card; the pair (*Kshosp*, *Kphosp*) is specific to one hospital; *IDApat|util* is dedicated to a single final user. Therefore, the risk of illicit disanonymization is considerably reduced. In the same way, the solution resists to dictionary attacks that could be led in different organizations: healthcare establishments, processing centers and final users.
- The combination of the suggested anonymization sequence (T_1, T_2, T_3) with access control mechanisms satisfies the non-inversibility requirement as well as the least privilege principle.
- It is possible to merge the data belonging to several establishments without compromising neither the security nor the 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 needs of other sectors (e.g., E-commerce, E-government, demographic studies, etc.).
- As smart cards are sufficiently tamper-resistant, their use seems suitable to keep secret the patient identifier. Moreover, smart cards are an adequate means to materialize the *patient consent*. Indeed, the patient medical data can appear in a database only if, by supplying his card, the patient gives his consent to exploit his medical data as a part of a project.

Besides, our solution regulates the medical data inversion. Let us take the example where the final user (i.e. researcher in rare or orphan diseases) discovers important information that necessitates re-identifying the patients. At first, it sends back results to the project center. The latter dispatches the results to the original hospitals participating to the concerned study (e.g., the orphan disease study).

Two cases can be identified:

- The original hospital has still the databases (or files) that allow establishing the link between the patient's identifiers, stay identifiers, and

medical data. In this case, the consulting physician performs the patient identification and informs him about the new research results.

- The hospital has deleted the nominative databases (for legal reasons or for security reasons); or the patient goes to a hospital participating to the project, but not the hospital where he was treated before. In these cases, by providing his medical data card (which implies that he gives explicitly his consent), it is possible to calculate $IDApat|Proj = H(IDproj \mid IDpat)$ and $IDAhosp(pat|Proj) = \{IDApat|Proj\}Kshosp$, and then, 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[4], would allow setting off an alarm. This alarm asks the patient if he wants to consult the results. Of course, if the knowledge of these results can harm the patient, it should contain a mention advising the patient to contact his consulting physician. The latter will inform him, in a suitable manner, about the results.

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

- The access to data has to be perfectly controlled; a well-defined security policy must be implemented by appropriate security mechanisms (hardware and/or software);
- The information system specification as well as the network architecture have to obey to a global security policy, and have to be adapted to needs;
- In some particular contexts, it is more efficient to completely separate identifier data from medical data.
- For repression or for deterring, it is recommended to control the purpose of use by calling for intrusion detection mechanisms; in particular, these mechanisms should detect malicious requests, illicit inferences, abuse of power, etc.

7 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 smartcards in this procedure responds to many security needs.

Although this solution is based on several successive anonymization steps, the cryptographic mechanisms that it uses are not expensive in terms of time and computation resources, and are compatible with current smartcard technology. Using Java Cards, we have implemented a prototype of this solution with a complete medical scenario, and we will soon be able to measure the performance and complexity of a real application.

8 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] U.S. Department of Health & Human Services, Update on the HIPAA Privacy and Security Final Rule, January 17, 2013.

[3] “Long-expected omnibus HIPAA rule implements significant privacy and security regulations for entities and business associates” Mayer Brown LLP, February 11, 2013.

[4] “HITECH Final Rule Results in Significant Changes to HIPAA Provisions” Faegre Baker Daniels, January 30, 2013.

[5] 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.

[6] Directive 95/46/CE of the European Parliament: “On the protection of individuals”; October 24, 1995.

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

[8] Loi 78-17 du 6 janvier 1978 relative à l’Informatique, aux fichiers et aux libertés, Journal officiel, pp. 227-231

[9] B. Claerhouta, G.J.E. DeMoor, "Privacy protection for clinical and genomic data: The

[4] This list is sent by the final user (i.e. the scientific researcher). It contains the anonymous identifiers with the results.

use of privacy-enhancing techniques in medicine", International Journal of Medical Informatics, Volume 74, Issues 2–4, March 2005, Pages 257–265, Elsevier.

[10] M. Hansen, P. Berlich, J. Camenisch, S. Clauß, A. Pfitzmann, M. Waidner, "Privacy-enhancing identity management", Information Security Technical Report, Volume 9, Issue 1, January–March 2004, Pages 35–44, Elseiver.

[11] M. Rahman, B. Carbunar, M. Banik, “Fit and Vulnerable: Attacks and Defenses for a Health Monitoring Device”, 13th Privacy Enhancing Technologies Symposium (PETS 2013), Bloomington, Indiana, USA, July 10 – 12, 2013, Springer LNCS.

[12] A. Abou El Kalam, Carlos Aguilar-Melchor, S. Berthold, J. Camenish, S. Clauß, Y. Deswarte, M. Kohlweiss, A. Panchenko, L. Pimenidis, M. Roy "Further Privacy Mechanisms", Chapter 18, in Digital Privacy: PRIME — Privacy and Identity Management for Europe, Jan Camenisch, Ronald Leenes & Dieter Sommer (Eds.), Springer, Lecture Notes in Computer Science (LNCS 6545), 2011, ISBN 978-3642190490.

[13] C. Quantin, H. Bouzelat, FA. Allaert, AM. Benhamiche, J. Faivre et L. Dusserre, “How to ensure data security of an epidemiological follow-up”, Medical Informatics 49 (1998).

[14] B Blobel, “Clinical Record Systems in Oncology. Experiences and Developments on Cancer Registers in Eastern Germany”, Personal Medical Information Security, Engineering and Ethics, ISBN 3-540-63244-1, 997.

[15] J.P. Jeanneret, D. Olivier, J. Chiffelle, “How to Protect Patient’s medical Secret in Official statistic”, Information Security Solutions Europe Conference, London, 2001.

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

[17] A. Menezes, P. C. Van Oorshot, S. A. Vanstone, “Handbook of Applied Cryptography”, 1997, CRC press, ISBN : 0849385237, pp. 780.