Service for the Pseudonymization of Electronic Healthcare Records Based on ISO/EN 13606 for the Secondary Use of Information

Roberto Somolinos, Adolfo Muñoz, M. Elena Hernando, *Senior Member, IEEE*, Mario Pascual, *Member, IEEE*, Jesús Cáceres, Ricardo Sánchez-de-Madariaga, Juan A. Fragua, Pablo Serrano, and Carlos H. Salvador, *Senior Member, IEEE*

***Abstract*—The availability of electronic health data favors scientific advance through the creation of repositories for secondary use. Data anonymization is a mandatory step to comply with current legislation. A service for the pseudonymization of electronic healthcare record (EHR) extracts aimed at facilitating the exchange of clinical information for secondary use in compliance with legislation on data protection is presented. According to ISO/TS 25237, pseudonymization is a particular type of anonymization. This tool performs the anonymizations by maintaining three quasi-identifiers (gender, date of birth, and place of residence) with a degree of specification selected by the user. The developed system is based on the ISO/EN 13606 norm using its characteristics specifically favorable for anonymization. The service is made up of two independent modules: the demographic server and the pseudonymizing module. The demographic server supports the permanent storage of the demographic entities and the management of the identifiers. The pseudonymizing module anonymizes the ISO/EN 13606 extracts. The pseudonymizing process consists of four phases: the storage of the demographic information included in the extract, the substitution of the identifiers, the elimination of the demographic information of the extract, and the elimination of key data in free-text fields. The described pseudonymizing system was used in three telemedicine research projects with satisfactory results. A problem was detected with the type of data in a demographic data field and a proposal for modification was prepared for the group in charge of the drawing up and revision of the ISO/EN 13606 norm.**

***Index Terms*—Electronic medical records, identification of persons, ISO standards, medical information systems, pseudonymization, telemedicine, web services.**

Manuscript received January 24, 2014; revised May 21, 2014 and August 14, 2014; accepted September 19, 2014. Date of publication September 26, 2014; date of current version November 3, 2015. This work was supported in part by Project PI08/1148, Project PI08/90330, Project PI12/01476, and Project PI12/01305 (coord. PI12/00508) from Fondo de Investigación Sanitaria (FIS) Plan Nacional de I+D+i and by Project CEN-20091043.

R. Somolinos and J. A. Fragua are with the Bioengineering and Telemedicine Laboratory, University Hospital Puerta de Hierro Majadahonda, Madrid 28222, Spain (e-mail: rsomolinos@idiphim.org; jafragua@idiphim.org).

A. Muñoz, M. Pascual, J. Cáceres, R. Sánchez-de-Madariaga, and C. H. Salvador are with the Telemedicine and Information Society Department, Health Institute "Carlos III" (ISCIII), Madrid 28029, Spain (e-mail: adolfo.munoz@isciii.es; mario.pascual@isciii.es; jcaceres@isciii.es; ricardo.sanchez@isciii.es; chsalvador@isciii.es).

M. E. Hernando is with the Bioengineering and Telemedicine Group, Polytechnic University of Madrid, Madrid 28040, Spain, and also with the CIBER-BBN: Networking Research Center for Bioengineering, Biomaterials, and Nanomedicine, Madrid 28029, Spain (e-mail: elena@gbt.tfo.upm.es).

P. Serrano is with the Fuenlabrada University Hospital, Madrid 28942, Spain (e-mail: pserrano.hflr@salud.madrid.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2014.2360546

I. Introduction

THE availability of open health data [1] for secondary use is fundamental for the advance in medical knowledge. The use of public datasets by researchers has repercussions on the acceleration of scientific advances as well as improvements in both the efficiency and efficacy of health processes [2], [3]. A requisite for the existence of public repositories of health data is to guarantee patient privacy by means of anonymization and de-identification techniques. The legislations of different countries establish that the secondary use of clinical data is only permitted if the exchanged information is previously anonymized to avoid the future association with its owners [4], [5].

This paper presents a pseudonymizing system for facilitating the exchange of data for secondary use in research projects. The system is designed and developed in accordance with the ISO/EN 13606 norm [6] and allows the total or partial anonymization of electronic healthcare record (EHR) extracts. The total anonymization eliminates all of the demographic references whilst the partial anonymization allows some of the demographic data (sex, date of birth, and place of residence) to be maintained at the precision selected by the users.

The anonymization process separates the clinical information and the associated demographic information. Basically, it consists of the elimination of the demographic information of the extract, prior to storage, and the substitution of all of the identifiers that might be associated with specific demographic entities. The management of the identifiers is of great importance in the pseudonymization process and is essential in permitting future associations of the demographic information.

In the communication between different systems, the clinical information is transmitted by following the mechanisms established in the norm, but instead of transmitting the complete EHR extract, the anonymized extract is transmitted. If the message is intercepted by external agents, the clinical information cannot be associated with a specific entity. The demographic information eliminated from the extracts is locally stored and may be recovered by means of consulting the identifiers that appear in the anonymized extract whenever the rights pertaining to access to the information are verified.

II. Background

Research in biomedical and health sciences, key instrument in the improvement in the quality of life of citizens, has changed

in recent years, both methodologically and conceptually, thanks to the appearance of new tools for the analysis of data [7]. Much has been legislated in recent years in this area, with special emphasis on that related to the access and use of personal data.

The European Union, by means of the 95/46/EC directive and the Article 29 Working Party, has established the mechanisms necessary to guarantee the protection of the individual as regards the handling and free circulation of personal data between its member states. It defines "personal data" as any information relating to an identified or identifiable natural person, and an "identifiable person" is one who can be identified, directly or indirectly (article 2a). The EC Data Protection Directive is not applied when the individual is not identifiable. The Article 29 Working Party has established "anonymous data" as any information related to a person who cannot be identified. According to ISO/TS 25237 "Health informatics – Pseudonymization" [8], "anonymization" is the process that removes the association between the identifying data set and the data subject and "pseudonymization" is a particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms. In accordance with the World Health Organization (WHO) guidelines, "proportional or reasonable anonymity exists when no reasonable means of identification of specific individuals is available." In January 2012, the European Commission proposed a comprehensive reform of the EU's 1995 data protection rules to strengthen online privacy rights and boost Europe's digital economy.

Spanish legislation follows the European 95/46/EC directive. According to the Spanish 14/2007 law on Biomedical research [9], (article 50, 2) data of a personal nature may only be used for research or teaching purposes when the interested party has expressly given his or her consent or when the said data has been previously anonymized, and (article 52, 3) the said data may only be preserved for research purposes in an anonymized format.

In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [10] is responsible, by means of the Privacy, Security and Patient Safety Rules, to protect the privacy of the clinical information and establish regulations to guarantee the security of the EHR. There is no legislative requirement to obtain patient consent to keep clinical information if the data are previously de-identified. The HIPAA Privacy Rule [11] is concerned with protection against identity disclosures and provides definitions and standards for the de-identification of clinical data. The HIPAA "Safe Harbor" defines 18 data elements called Protected Health Information (PHI) that must be removed to consider that clinical data are de-identified.

Therefore, those scenarios in which an information system has to send clinical information to the exterior for secondary use need this information to be anonymized previously. For this reason, it is proposed to design and develop an anonymizing system.

The anonymization of sensitive information is a broadly addressed problem, and there is an amount of solutions including pattern matching and machine learning methods [12], [13]. Quasi-identifiers are defined as those variables representing environment data that could be used to re-identify a person. The equivalence classes are the sets of the records having the same values from a set of selected quasi-identifiers. In *k*-anonymity [14], the minimum size of all established equivalence classes is defined as *k*. This means that for any register, there are at least $k-1$ other registers with the same values of the quasi-identifiers. In order to guarantee a low re-identification risk, a minimum value of *k* must be guaranteed. *k*-Anonymity prevents identity disclosures. There are other later models such as l-diversity and t-closeness [15] studying the probability and distribution of the sensitive attributes and protecting against attribute disclosures.

Many investigations need to know certain personal data from patients in order to produce significant results; as a consequence anonymization cannot be total and there is a risk of re-identifying participant patients through present quasi-identifiers. The most frequent and important quasi-identifiers for secondary use are gender, date of birth, and place of residence. Following a study by Sweeney [16], 87% of the population in the United States could be uniquely identified by these three quasi-identifiers: gender, date of birth, and their five-digit ZIP code. The most extended solution in order to reduce the re-identification probability is to group quasi-identifiers so that the number of equivalence classes diminishes and their size grows as does the value of *k*. The quasi-identifiers at the extremes must be truncated in order to avoid re-identifications through unusual values.

III. Methods

The ISO/EN 13606 norm, drawn up by the European Committee for Standardization (CEN) [17], has as its main objective the standardization of EHR transfers, or part of them, in a semantically operable manner. It is based on the following paradigms: separation of responsibilities (division of a complex problem into several simpler subproblems), separation of points of view (definition of five points of view of the distributed systems: business, information, computation, engineering, and technology [18]), together with the separation of information and knowledge. In accordance with the last of the paradigms, this standard separates the information from the knowledge right from its design stage, basing it on a double model [19]. There is a similar philosophy in the solutions proposed by other organizations such as HL7 [20] and openEHR [21], there being agreements between these organizations to reach common interoperability solutions, such as the CIMI initiative [22].

The double model of the norm consists of the reference model (information model) that defines the structures necessary to organize the information and the archetype model (knowledge model), which represents the domain of the knowledge formally modeling the concepts. Both models are mutually complemented and are necessary to achieve the interoperability of the clinical information.

The norm describes a reference model (see Fig. 1) that provides the classes necessary to represent the clinical information and its context. The extract (*EHR_EXTRACT* class) is the basic information transmission unit. The reference model includes

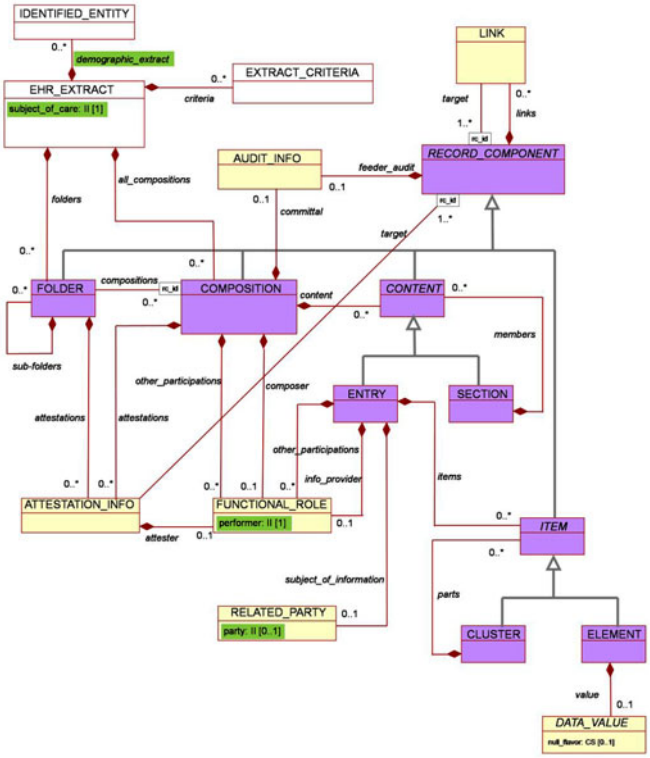

Fig 1. Reference model of the ISO/EN 13606 norm.

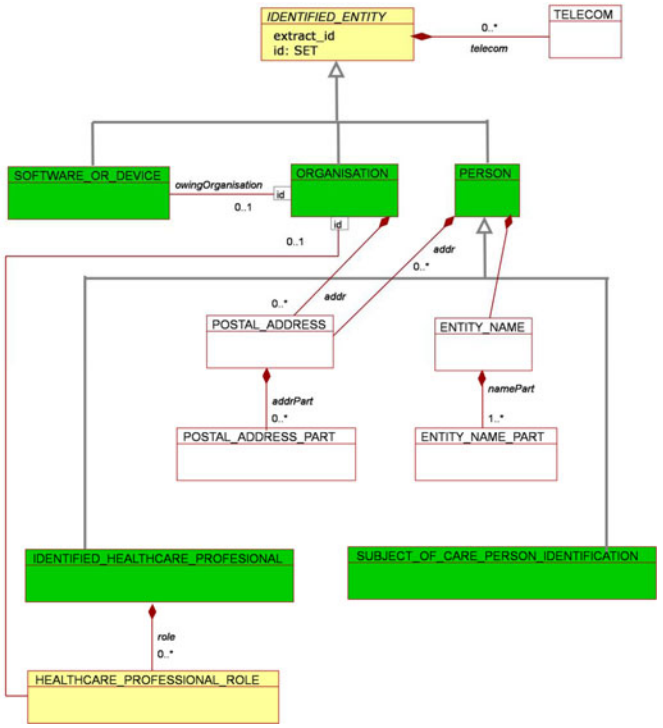

Fig. 2. Demographic package of the ISO/EN 13606 norm.

a separate package for the demographic information of all of those actors that intervene in the EHR (patients, health staff, organizations, devices, etc.). The *EHR_EXTRACT* class includes the *demographic_extract* field in which the data on the participating demographic entities are stored in accordance with the types of demographic package. The rest of the fields of the *EHR_EXTRACT* class are used to represent the clinical information and its context. This data separation is very useful at the time of anonymizing the information, since it allows the demographic entities to be represented in the clinical part of the extracts only by means of an identification code, as occurs in the *subject_of_care* field of the *EHR_EXTRACT class*, the *party* field *of the RELATED_PARTY* class and the *performer* field of the *FUNCTIONAL_ROLE* class. In this way, the management of the identifiers of the demographic entities and the elimination of the tracing of the subjects of the clinical data is facilitated. Fig. 1 shows the reference model of the ISO/EN 13606 norm, emphasizing its link with the demographic package and the identifying fields used to represent the demographic entities.

The demographic package of the ISO/EN13606 norm and the relationships existing between its classes are shown in Fig. 2. The main class (*IDENTIFIED_ENTITY*) of the demographic package is an abstract class that encompasses all of the classes of demographic entities. *IDENTIFIED_ENTITY* is implemented by the rest of the specific classes of the package that represent the different types of entity: *SOFTWARE_OR_DEVICE*, *ORGANISATION*, *PERSON*, *IDENTIFIED_HEALTHCARE_PROFESSIONAL*, and *SUBJECT_OF_CARE_PERSON_IDENTIFICATION*. These classes inherit the fields of their mother class, which will be common in all of the classes of demographic entities. These fields are:

- *extract_id:* Type *II* field (*InstanceIdentifier*), unique identifier used to represent this demographic entity within the extract.
- *id:* Series of type *II* identifiers from which this demographic entity may be referenced (national identify, social security, hospital numbers).

The type of data *II* (*InstanceIdentifier*) is used to represent identifying objects. Class *II* contains six fields, but the most important ones are *root* and *extension*, since two *II* objects are considered equals if and only if their values of the *root* and *extension* fields are the same, that is, both objects would identify the same instance. The *root* field is a unique identifier that guarantees the overall uniqueness of the type *II* objects. It makes up a type of "names space," which, by means of a code assigned to an entity ensures that all of the *II* objects that are generated under its supervision are unique. The *extension* field is a chain of characters that makes up a unique identifier within the name space specified by *root*.

As a result of previous projects in the field of standardizing the transfer of EHR, our group has developed an EHR server in accordance with the ISO/EN13606 norm [23]. Libraries were generated to represent the reference model, the demographic package and the types of data of the ISO/EN13606 standard. The server was developed using Java as its programming language, MySQL [24] for the databases, XML [25] as the mark-up language, XML schemas [26] to design the structure of the data, JPA libraries [27] for the permanent storage, JAXB libraries [28] for the automatic generation of Java classes, and Web Services

were used as communication technologies [29] implemented by means of the Axis2 tool and deployed on an Apache Tomcat applications server [30]. This previous work supports the current design and development of the pseudonymizing system.

IV. Results

The pseudonymizing system was designed, developed, and tested in accordance with the ISO/EN 13606 norm. The main function of this system is to eliminate links between the demographic data and the ISO/EN13606 clinical information extracts that are sent to other entities for secondary use. Its main characteristics are as follows:

- It is an independent module integrated within the information generator and emitter of EHR extracts system in such a way that nonanonymized information is never sent out of the system of origin in compliance with the relevant legislation.
- In total anonymizations, it eliminates all of the explicit demographic information of the extracts.
- In partial anonymizations, it keeps certain quasi-identifiers with a degree of specification configurable by the user.
- It eliminates the identifiers and quasi-identifiers (names, addresses, and dates of birth) that can appear in any free-text field.
- The demographic information is not lost, as it is registered and stored correctly, allowing both its later recovery by the duly authorized entities by means of future consultations, and the maintenance of the coherence of the provided identifiers.
- The same identifier always corresponds to a specific entity in extracts belonging to the same project (under the same "names space"), maintaining the coherence between the assigned identifiers. That is, if two extracts from the same project refer to the same demographic entity, the identifier used is the same.
- All references to the participating entities within the EHR extract are eliminated by means of type *II* identifiers. For this reason a mechanism to create, substitute and manage identifiers is enabled to make it impossible for external entities to establish links between the new identifiers and their corresponding demographic entities.

The clients of this pseudonymization tool are required to perform a previous population study of their samples in order to select the degree of specificity of the adequate quasi-identifiers guaranteeing the required k value for k-anonymization. The system provides the tools to perform anonymization in accordance with these parameters in a systematic way. Since this tool is aimed to help very different natured projects with very different attributes, it is not possible to implement more powerful models such as l-diversity and t-closeness, which are based on the values of the sensitive attributes.

The system is made up of two modules: a demographic server and a pseudonymizing module. Both modules use web services to offer access to its clients by means of a series of public functions. The demographic server can work in a totally independent way with clients who wish to save or recover the demographic information of certain entities. However, the pseudonymizer always works collaboratively with an associated demographic server, acting as a client of the functions offered by it.

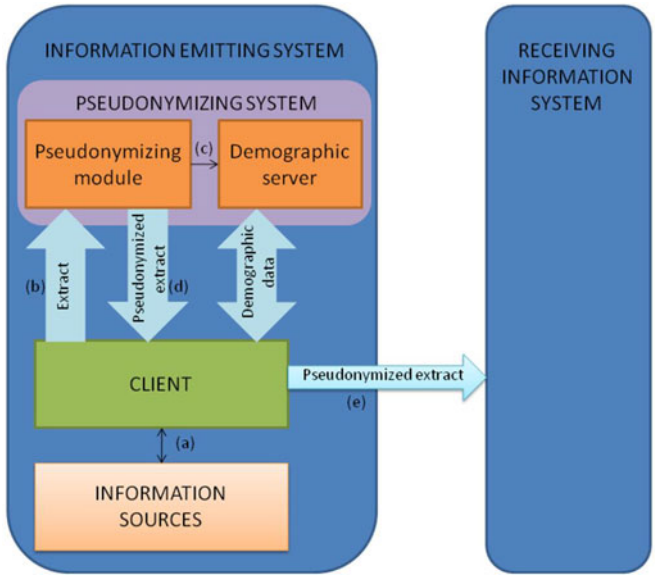

Fig. 3. Workflow in the sending of pseudonymized information.

The work flow in the sending of pseudonymized information between two heterogeneous systems is detailed in Fig. 3. The pseudonymizing system is integrated within the structure of the information emitting system. The clients make up ISO/EN13606 extracts containing the information they wish to send from extracts already generated by different sources [see Fig. 3(a)]. The extracts are passed on to the pseudonymizer [see Fig. 3(b)] for its treatment before being sent to the remote reception system [see Fig. 3(e)]. The suppressed demographic information is stored in the associated demographic server [see Fig. 3(c)]. If the clients within the emitting system are duly authorized, they could interact directly with the demographic server to register or recover demographic data.

The demographic server has two main objectives: the permanent storage of the demographic entities and to facilitate the demographic entity identifiers management.

- The permanent storage is carried out by means of two functions (*registerIdentifiedEntity* and *recoverIdentifiedEntity*), which allow its clients to save demographic entities in the storage system and recover them by means of their identifiers. The said identifiers are of type *II* and are defined by the values of their *root* and *extension* fields. The demographic entities dealt with by the server are transmitted as *IDENTIFIED_ENTITY* objects.
- The demographic server also has several functions that serve as tools for the management of the identifiers on behalf of its "pseudonymizer" clients. These functions allow to know whether a specific demographic entity already exists in the server by means of any of its type *II* identifiers (*existII* function), find out the value of the *extension* field of two identifiers that refer to the same demographic entity (*equivalentExtension* function) and update the data

bases of the server in order to add a new type *II* identifier to the list of a specific demographic entity (*updateSetId* function).

The pseudonymizing module is in charge of pseudonymizing the ISO/EN13606 extracts by generating new identifiers starting from a value of the given *root* field, called *rootProject*. This *rootProject* value should be common for all of the anonymizations within the same secondary use project and represents a names space within which the anonymous identifiers are generated for all of the entities. When receiving an extract, the pseudonymizer sends its demographic information to the demographic server for its storage and generates a new extract where the relevant type *II* identifiers are substituted by others within the common names space. The new extracts contain the same clinical information as the initial ones keeping just the demographic information selected by the client from among the following options:

- Gender: a) removed, and b) included.
- Date of birth: a) removed, b) groups of ten years, c) groups of five years, d) year, e) month, and f) day.
- Place of residence: a) removed, b) country, c) state, d) city, e) postal or zip code, and f) all included.

A detailed description of the interactions among the clients, the pseudonymizing module, and the demographic server is shown in points (a) through (f) in the text and in Figs. 4 and 5.

The extract pseudonymization begins with a call to the *anonymizeExtract* function (a) and is made up of the following phases:

1) Storage of the demographic information included in the extract. The ISO/EN 13606 extract has a non-obligatory field called *demographic_extract* in which the demographic data of entities related to the extract are included. For each of these entities it is checked whether it is already stored (b) in the associated demographic server (by means of its series of identifiers). In the affirmative case, the series of identifiers are updated (c), if necessary, to enter the new data of the extract. If this is not the case, the complete entity is registered and stored (d) in the demographic server.
2) Substitution of the entities' identifiers of the extract. In the clinical part of the extract there are several fields linked by type *II* identifiers that refer to demographic entities that intervene in the extract. Although the identifiers in themselves do not contain demographic information, they must be substituted since the external agents would already know to which entity each identifier refers. The clearest case is the subject of attention, specifically the *subject_of_care* field of the *EHR_EXTRACT* class. But there are other, less obvious, fields that must equally be anonymized, such as the *party* field of the *RELATED_PARTY* class and the *performer* field of the *FUNCTIONAL_ROLE* class (see Fig. 1). The new type *II* identifiers will have *rootProject* as the value of its *root* field. The value of its *extension* field is assigned in such a way as to ensure that there are no replicated identifiers (e) and the stored demographic entities are updated in the server (f) with the new assigned identifiers. If any entity has already been handled in the same project and therefore already has an identifier with a *rootProject* value in its *root* field, the said identifier will be used to include it in the pseudonymized extract thus keeping the coherence of the information.
3) Suppression of the demographic information included in the extract. All of the data included in the *demographic_extract* field of the EHR extract are eliminated, except those related to the gender (*degreeG*), date of birth

```
STEP 1: Storage of the demographic information included in the extract

listIdentifiedEntity = ehrExtract.getDemographicExtract()
for (ie from listIdentifiedEntity) {
   registeredId = null
   listId = ie.getId()
   for (id from listId) {
      if (existII(id)) (b) then registeredId=id
   }
   if (registeredId != null) then {
      listId2 = ie.getId()
      for (id2 from listId2) {
         if (!(existII(id2))) then updateId (ie, id2) (c)
      }
   }
   else {
      register(ie) (d)
   }
}

STEP 2: Substitution of the identifiers of the entities of the extract

if (subject_of_care != null) {
   oldExtensionSOC = subject_of_care.extension
   subject_of_care = anonymizeII(subject_of_care, rootProject) (e,f)
   newExtensionSOC = subject_of_care.extension
}
if (lookForParty) {
   party = anonymizeII(party, rootProject) (e,f)
}
if (lookForPerformer) {
   performer = anonymizeII(performer, rootProject) (e,f)
}

STEP 3: Supression of demographic information included in the extract

SOC = new SUBJECTOFCAREPERSONIDENTIFICATION()
SOC.administrativeGenderCode = selectGender(degreeG,
                               subject_of_care.administrativeGenderCode)
SOC.birthTime = selectBirthDate(degreeB, subject_of_care.birthTime)
SOC.addr.postalCode = selectPC(degreeA, subject_of_care.addr.postalCode)
for (pap from SOC.addr.addrPart) {
    pap.addressLine = selectAL(degreeA, pap.addressLineType)
}
demographic_extract=null
demographic_extract.add(SOC)

STEP 4: Removal of key data in free-text fields

if (find(oldExtensionSOC)) then change(oldExtensionSOC, newExtensionSOC)
if (find(SOC.name.namePart.entityPartName)) then remove
if (find(SOC.addr.addrPart.addresLine)) then remove
if (find(SOC.addr.postalCode)) then remove
if (find(SOC.birthTime)) then remove
```

Fig. 4. Pseudocode of the pseudonymization process.

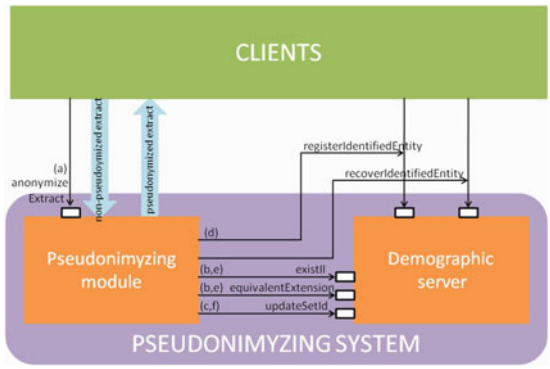

Fig. 5. Interaction between modules and clients in the pseudonymization process.

(*degreeB*) and place of residence (*degreeA*) of the entity of the subject of care, which are indicated with the degree of specification selected by the client. These data are useful for secondary uses as statistical and research studies and the client is responsible for choosing the sufficiently general degrees to ensure that the risk of re-identification of the patients is low.

4) Elimination of key data in the free text field. Although the ISO/EN 13606 norm is designed to be able to include any type of data in a structured way, it is common for many extracts to include data that allow access to the demographic information in free text fields. For this reason, a mechanism was established to detect and correct these types of cases. It consists of the search in all of the already anonymized extract (all text fields, i.e., *EXTRACT_CRITERIA/other_constraints, RELATED_PARTY/relationship*, etc.) for key data, such as the identifiers (*extension* fields), names (*entityPartName*), addresses (*postalCode* and *addressLine*) and dates of birth (*birthTime*) of the participating demographic entities. If any result is found in the search, it goes on to eliminate the said key datum and, if it is an identifier, substitute it for its pseudonym.

Fig. 4 shows, by means of a pseudocode, the specific actions to be carried out in the pseudonymization process.

Fig. 5 shows the functions accessible by means of the web services of the two modules that make up the pseudonymizing system, as well as the pseudonymizing process in detail.

Different examples of the anonymization of extracts are described in detail in the appendix (online version), which include the most representative cases in the flow of anonymization tasks and how the different functions are used depending on the data included in the extract and the information already stored in the demographic server.

V. Discussion

The described pseudonymizing system has been integrated into the following Telemedicine projects:

1) In the CAMAMA project, carried out in conjunction with the Hospital de Fuenlabrada (Madrid) and the Hospital Clinic (Barcelona), which studied the automation of the sending of clinical information between producers (hospitals) and consumers (biobanks, case registers, and other research groups). In principle, the objective was to cover only cases of cancer, but it was finally extended to cover all of the patients of the Fuenlabrada hospital. This still active project aims to reach the figure of 200 000 summarized clinical health records exchanged by means of pseudonymized extracts.
2) In the OBESITY project, the monitoring and control of obesity data in the primary attention of the 206 patients included were sent in anonymized extracts to the central node for its later secondary use.
3) In the REHABILITA project, pseudonymized data originating from the rehabilitation sessions of patients in specific scenarios were exchanged to cover a wide range of frameworks using peer-to-peer architectures for the exchange of clinical information between nodes.

In the three projects, the extracts were pseudonymized correctly, making it impossible to establish links between the clinical data and its owners by means of the identifiers existing in the anonymized extracts, although there is a risk of re-identification by means of the quasi-identifiers present. In order to check the correct functioning of the system, several tests on the pseudonymized extracts are performed; these tests perform manual search of key data that could cause re-identification. In the case of the OBESITY and REHABILITA projects, in which the number of extracts is low, all of them have been checked. In the case of CAMAMA, a random sample (approximately 10% of the received extracts) is analyzed periodically. So far no problem was found in the pseudonymized information.

In those cases in which the emitting system also wanted to recover demographic information from the identifiers of the pseudonymized extract, it was achieved successfully. In the aforementioned tests, it was also confirmed that the same identifier was assigned in the anonymizations of the same entity in the same project, thus maintaining the coherence for secondary use.

The selection of the granularity of the quasi-identifiers has been different in each project depending on their own characteristics. The researchers in these projects are responsible for selecting a configuration that guarantees a low risk of re-identification. The configurations chosen are as follows:

- CAMAMA: Gender: included, Date of birth: year and Place of residence: removed.
- OBESITY: Gender: included, Date of birth: groups of five years and Place of residence: postal or zip code.
- REHABILITA: Gender: included, Date of birth: groups of five years and Place of residence: state.

The CAMAMA project establishes its equivalence classes from the gender and the year of birth of the patients. A grouping together has been carried out by age in all of the registers for those people over 80 years' old so as to reduce the probability of re-identification. At the time of writing this paper, there are 30 000 registers available, although the aim is to reach the entire population of Fuenlabrada, which has more than 200 000 inhabitants. By means of the population pyramid of Fuenlabrada [31], it has been possible to approximate the size of the equivalence classes. The most numerous of the groups is that of men between 30 and 34 years old with 9770 elements and the least numerous is that of men between 75 and 79 years old with 1100 integrants. Assuming a uniform distribution in these five years, the equivalence classes obtained with the least number of integrants contains 220 elements, for which the value of k is 220 with these anonymization parameters, which guarantees an acceptable risk of re-identification for this project.

During the design of the pseudonymizing system a problem was detected in the model of the demographic package of the ISO/EN 13606 norm. The *birthTime* field of the *SUBJECT_OF_CARE_PERSON_IDENTIFICATION* class is of the *TS* (*TimeStamp*) type. This implies that the minimum degree of specification of the dates, which allows this field to be stored is the year, making it impossible to use it to indicate dates in wider

Fig. 6. Archetype of the composition "Other demographic data."

ranges. As a solution, a mechanism has been enabled outside the demographic package to represent wider ranges of dates using a *COMPOSITION* object along the lines of the archetype shown in Fig. 6. However, the definitive solution must undergo the modification of the reference model of the norm in its revisions.

The pseudonymizing system is designed to be integrated into the original information system. In this fashion, information will never go unanonymized outward. It also trusts in the mechanisms of the original information system to detect inconsistencies in the patients' data. We are working in a mechanism that permits to update the demographic server data from inconsistencies detected and communicated by the original system.

As an alternative scenario, the pseudonymizing system could be deployed independently of any information system. It would be a Trusted Third Party (according to ISO/TS 25237) for the information systems willing to use its service. In this way, service could be supplied to projects having their data in separate information systems. In the case of this configuration, it is essential that the communication between the information systems and the pseudonymizer be safe, since nonanonymized data are being transmitted.

One detected limitation of the current system is that multimedia data in the extracts can contain patients' information (i.e., DICOM images, *attested_view* screenshots) and they are not anonymized. Other detected limitation is the apparition of other identifiers (for instance pseudonyms) occurring in textual fields not included in the *demographic_extract* or not registered in the demographic server that are not detected. To solve these limitations, the original information system should make sure that these circumstances will not happen before anonymization.

VI. Conclusion

The pseudonymization of the EHR extracts thwarts the identification of the clinical data owners. In this way, it facilitates compliance with current data protection legislation in the exchange of clinical information for secondary use. Different research projects present different necessities as regards the de-identification of their patients, there being a balance between the data present for the extraction of results and the risk of re-identification. The presented tool allows the de-identification in different scenarios to take place, the researchers being responsible for the choice of the most suitable parameters given the characteristics of each project.

The reference model of the ISO/EN 13606 clearly separates the clinical and the demographic information in such a way that if there were references to demographic entities within the clinical information, they could only be implemented by means of identifiers. On the other hand, the demographic entities might have more than one identifier to be referenced, and this is indicated in the field proper to the demographic entity. These characteristics, proper to the norm, facilitate the management of the identifiers and the anonymization of the extracts. Because of this, the norm brings together the ideal characteristics to be the basis of a pseudonymizing system. As a consequence of this study, a modification is going to be proposed to the implementation and revision group of the ISO/EN 13606 norm, in which some of the authors participate so that the demographic package of the reference model is able to register an interval in the date of birth.

When carrying out the pseudonymization of the extracts, the demographic information of all of the entities participating in the demographic server is also saved. The said information will always be available, both for the project that generated it and for any other use within the information system into which the pseudonymizing system is integrated. Any suitable application would be able to request the said data from the demographic server, and this may be shown by previously proving that the applicant has the permission necessary to access the said information.

The pseudonymizing system has been integrated and used in active telemedicine projects of our work group. It has gone on to form part of the platform of services offered by our research unit [32] together with other services such as the storage and recovery of ISO/EN13606 extracts, the validation of extracts, archetype editor and server [33], and randomization service.

The repository of anonymized information obtained from the aforementioned research projects, has permitted a promising line of research to commence based on KnowledgeDiscovery in Databases (KDD) and Data Mining (Association Rules) techniques [34] for the extraction of clinical knowledge.

References

[1] N. M. ÓBoyle, R. Guha, E. L. Willighagen, S. E. Adams, J. Alvarsson, J. C. Bradley, I. V. Filippov, R. M. Hanson, M. D. Hanwell, G. R. Hutchison, C. A. James, N. Jeliazkova, A. S. Lang, K. M. Langner, D. C. Lonie, D. M. Lowe, J. Pansanel, D. Pavlov, O. Spjuth, C. Steinbeck, A. L. Tenderholt, K. J. Theisen, and P. Murray-Rust, "Open data, open source and open standards in chemistry: The blue obelisk five years on," *J. Cheminform*, vol. 3, no. 1, p. 37, Oct. 2011.

[2] S. E. Fienberg, "Sharing statistical data in the biomedical and health sciences: Ethical, institutional, legal, and professional dimensions," *Annu. Rev. Public Health*, vol. 15, pp. 1–18, 1994.

[3] H. A. Piwowar, R. S. Day, and D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate," *PLoS One*, vol. 2, no. 3, p. e308, 2007.

[4] B. S. Elger, J. Iavindrasana, I. L. Lo, H. Müller, N. Roduit, P. Summers and J. Wright, "Strategies for health data exchange for secondary, cross-institutional clinical research," *Comput Methods Programs Biomed.*, vol. 99, no. 3, pp. 230–251, Sep. 2010.
[5] C. Weng, P. Appelbaum, G. Hripcsak, I. Kronish, L. Busacca, K. W. Davidson, and J. T. Bigger, "Using EHRs to integrate research with patient care: Promises and challenges," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 684–687, Sep. 2012.
[6] *Electronic Health Record Communication Part 1: Reference Model*, ISO 13606-1, 2008.
[7] P. Libin, G. Beheydt, K. Deforche, S. Imbrechts, F. Ferreira, K. Van Laethem, K. Theys, A. P. Carvalho, J. Cavaco-Silva, G. Lapadula, C. Torti, M. Assel, S. Wesner, J. Snoeck, J. Ruelle, A. De Bel, P. Lacor, P. De Munter, , E. Van Wijngaerden, M. Zazzi, R. Kaiser, A. Ayouba, M. Peeters, T. de Oliveira, L. C. Alcantara, Z. Grossman, P. Sloot, D. Otelea, S. Paraschiv, C. Boucher, R. J. Camacho, and A. M. Vandamme, "RegaDB: Community-driven data management and analysis for infectious diseases," *Bioinformatics*, vol. 29, no. 11, pp. 1477–1480, Jun. 2013.
[8] *Health Informatics—Pseudonymization*, ISO/TS 25237, 2008.
[9] *Spanish Law 14/2007 on biomedical research*. (2007). [Online]. Available: http://www.boe.es/boe/dias/2007/07/04/pdfs/A28826–28848.pdf
[10] *Health Information Privacy* [Online]. Available: http://www.hhs.gov/ocr/privacy
[11] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 169–177, Mar. 2010.
[12] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: A review of recent research," *BMC Med. Res. Methodol.*, vol. 10, p. 70, 2010.
[13] K. El Emam, L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard and J. Gluck, "De-identification methods for open health data: The case of the heritage health prize claims dataset," *J. Med. Internet Res.*, vol. 14, no. 1, p. e33, 2012.
[14] K. El Emam and F. K. Dankar, "Protecting privacy using *k*-anonymity," *J. Amer. Med. Inform. Assoc.*, vol. 15, no. 5, pp. 627–637, Sep. 2008.
[15] S. Yoo, M. Shin, and D. H. Lee, "An approach to reducing information loss and achieving diversity of sensitive attributes in *k*-anonymity methods," *Interact J. Med. Res.*, vol. 1, no. 2, p. e14, 2012.
[16] L. Sweeney. (2000). Simple demographics often identify people uniquely. Carnegie Mellon University, Pittsburgh, PA, USA, Data Privacy Working Paper 3 [Online]. Available: http://dataprivacylab.org/projects/identifiability/paper1.pdf
[17] CEN. European Committee for Standardization. [Online]. Available: http://www.cen.eu
[18] *Information Technology—Open Distributed Processing—Reference Model: Architecture*, ISO/IEC 10746-3, 2009.
[19] T. Beale. *Archetypes: Constraints-based domain models for future-proof information systems*. (2002). [Online]. Available: http://www.openehr.org/files/resources/publications/archetypes/archetypes_beale_oopsla_2002.pdf
[20] Health Level Seven International. [Online]. Available: http://www.hl7.org
[21] OpenEHR. [Online]. Available: http://www.openehr.org
[22] CIMI Wiki. *Clinical information modeling initiative* [Online]. Available: http://informatics.mayo.edu/CIMI
[23] A. Munoz, R. Somolinos, M. Pascual, J. A. Fragua, M. A. González, J. L. Monteagudo, and C. H. Salvador "Proof-of-concept design and development of an EN13606-based electronic health care record service" *J. Amer. Med. Inform. Assoc.*, vol. 14, no. 1, pp. 118–129, Jan. 2007.
[24] *MySQL* [Online]. Available: http://www.mysql.com
[25] *Extensible Markup Language (XML)* [Online]. Available: http://www.w3.org/XML
[26] *XML schema* [Online]. Available: http://www.w3.org/XML/Schema
[27] Java Persistence API. [Online]. Available: http://en.wikipedia.org/wiki/Java_Persistence_API
[28] JAXB. [Online]. Available: http://jaxb.java.net
[29] Web Services Activity. [Online]. Available: http://www.w3.org/2002/ws
[30] Apache Tomcat. [Online]. Available: http://tomcat.apache.org
[31] *Population structure of Fuenlabrada*. (2013). [Online]. Available: http://ayto-fuenlabrada.es/recursos/doc/SC/Estadisticas_y_territorio/36781_1111112013133426.pdf
[32] Interoperability Services Platform. (2013). [Online]. Available: https://hce13606.telemedicina.isciii.es:8443/interServer
[33] R. Sanchez-de-Madariaga, A. Munoz, J. Caceres, R. Somolinos, M. Pascual, I. Martínez, C. H. Salvador and J. L. Monteagudo, "ccML, A new mark-up language to improve ISO/EN 13606-based electronic health record extracts practical edition," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 2, pp. 298–304, Mar. 2013.
[34] I. H. Witten, E. Frank, and A. H. Mark, *Data mining. Practical machine learning tools and techniques*, 3rd ed. Oxford, U.K.: Elsevier, 2011.

Authors' photographs and biographies not available at the time of publication.