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, Mario Pascual, Jesús Cáceres, Ricardo Sánchez-de-Madariaga, Juan A Fragua, Pablo Serrano, Carlos H Salvador

IEEE J Biomed Health Inform. 2015 Nov;19(6):1937-44.

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, 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 the health data favors scientific advance. The creation of repositories for secondary use is dependent on the anonymization of their data 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 of key data in free-text fields. The pseudonymizing system described has been used in three Telemedicine research projects with satisfactory results. A problem has been detected with the type of data in a demographic data field and a proposal for modification has been 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, telemedicine, web services**

Manual received ****** **, ****; revised ***** **, ****. Current version published ****** **, ****. This work was supported in part by projects PI08/1148, PI08/90330 and 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, Spain (e-mail: {rsomolinos;jafragua}@idiphim.org).

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

M.E. Hernando is with the Bioengineering and Telemedicine Group, Polytechnic University of Madrid, Madrid, Spain (e-mail: elena@gbt.tfo.upm.es).

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

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 in secondary use scenarios by means of anonymization and de-identification techniques. The legislations of different countries establish that the exchange of clinical data for secondary use is only permitted if the information exchanged is anonymized previously. It is therefore necessary for information transferred for secondary use not to be associated with its owners [4].

With the objective of facilitating the exchange of data for secondary use in research projects [5], a pseudonymizing system has been designed and developed in accordance with the ISO/EN 13606 norm [6]. This tool allows the total or partial anonymization of electronic healthcare records (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, place of residence) to be maintained at the precision selected by the users of the tool.

The anonymization of the EHR extracts allows to separate 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 present in the extract 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.

ln 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 not lost, but is stored correctly 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, "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 [8], (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) [9] 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 HIPPA Privacy Rule [10] is concerned with protection against identity disclosures and provides definitions and standards for the de-identification of clinical data. The HIPPA "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 [11-12]. 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 [13], 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-anonimity prevents identity disclosures. There are other later models such as l-diversity and t-closeness [14] 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 [15], 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. Method

The ISO/EN 13606 norm, drawn up by the European Committee for Standardization (CEN) [16], 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 sub-problems), separation of points of view (definition of five points of view of the distributed systems: business, information, computation, engineering and technology [17]), 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 [18]. There is a similar philosophy in the solutions proposed by other organizations such as HL7 [19] and openEHR [20], there being agreements between these organizations to reach common interoperability solutions, such as the CIMI initiative [21].

The double model of the norm consists of the reference model (information model) which 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 (Fig. 1) which 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 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 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.

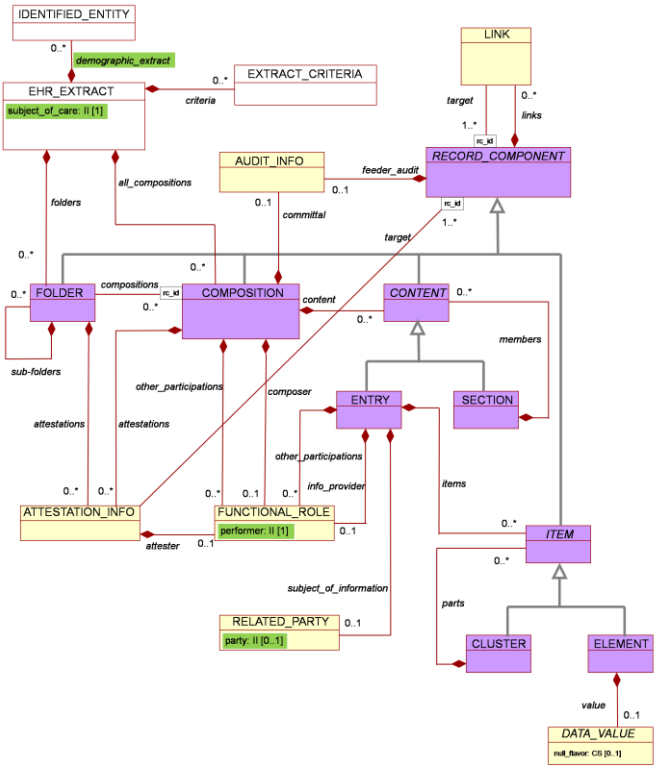

Fig 1. Reference model of the ISO/EN 13606 norm

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)

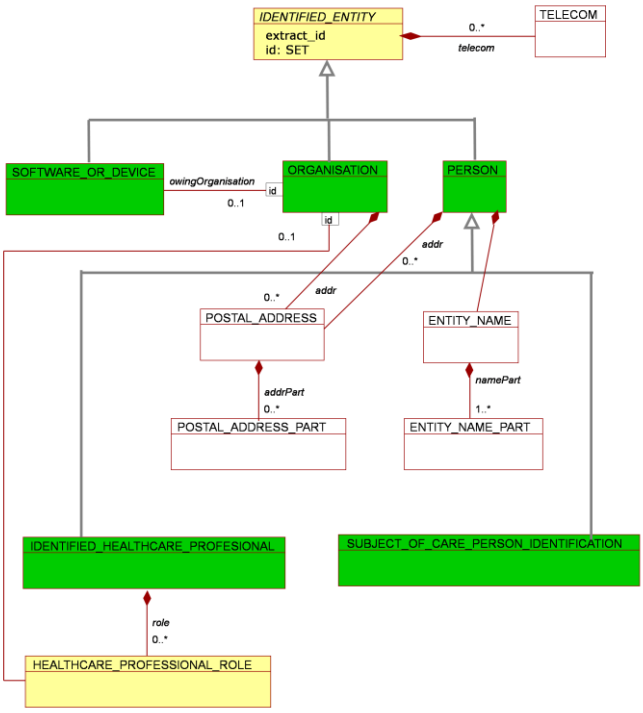

Fig. 2. Demographic package of the ISO/EN 13606 norm

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 they define which *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 this work of standardizing the transfer of EHR, our group has developed an EHR server in accordance with the ISO/EN13606 norm [22]. 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 [23] for the data bases, XML [24] as the mark-up language, XML Schemas [25] to

design the structure of the data, JPA libraries [26] for the permanent storage, JAXB libraries [27] for the automatic generation of Java classes and Web Services were used as communication technologies [28] implemented by means of the Axis2 tool and deployed on an Apache Tomcat applications server [29]. This previous work supports the design and development of the pseudonymizing system, based on the same technologies as the EHR server.

IV. Results

A pseudonymizing system has been 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 which are sent to other entities for secondary use. Its main characteristics are as follows:

- It is an independent and integrated module within the information generator and emitter of EHR extracts system in such a way that non-anonymized information is never sent out of the system of origin in compliance with the relevant legislation
- It eliminates all of the explicit demographic information of the extracts in total anonymizations
- It keeps certain quasi-identifiers with a degree of specification configurable by the user in partial anonymizations
- It eliminates the identifiers and quasi-identifiers (names, addresses, dates of birth) that can appear in any free-text field
- The demographic information is not lost, but 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 system under development is a tool to achieve clinical data pseudonymization for ulterior secondary use. The clients of this 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-adversity 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.

The pseudonymizing system is integrated within the structure of the information emitting system. The clients of this entity make up extracts with the information that they wish to send, from extracts already generated by different sources (a), be they their own or external. The extracts made up in this way are passed on to the pseudonymizer (b) for its treatment before being sent by the network to the reception system (e). The demographic information previously contained in the extracts is not lost, but stored in the associated demographic server (c). The clients, within the emitting system, are also able to interact directly with the demographic server if they are duly authorized, be it to save or recover demographic data. The work flow in the sending of pseudonymized information between two heterogeneous systems is detailed in Fig. 3.

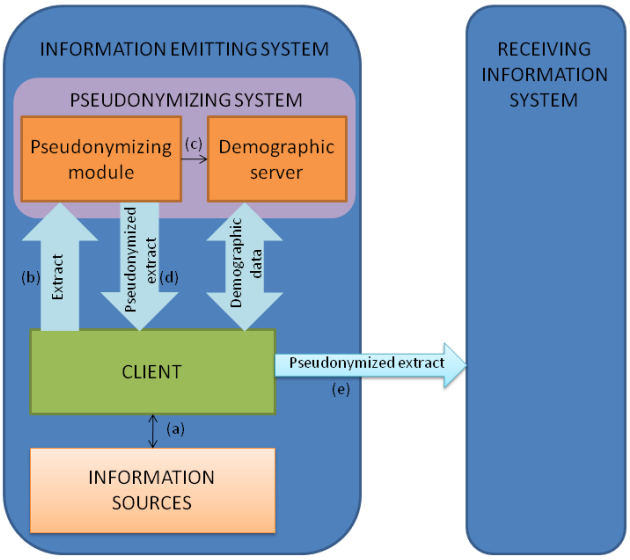

Fig. 3. Workflow in the sending of pseudonymized information

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 it to be known whether there is a specific demographic entity or not 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 from a value of the given *root* field, called *rootProject*. This *rootProject* value, common for all of the anonymizations carried out within the same project or application, represents a names space within which the anonymous identifiers are generated for all of the entities. The pseudonymizer sends all of the demographic information of the extracts to the associated demographic server for its storage, manages all of the relevant type *II* identifiers that appear in the extracts to be substituted by others within the common names space. The pseudonymizer generates new extracts with the same clinical information as the initial ones, with new identifiers which have no relationship with the previous ones and only with certain demographic information relative to the patient. The pseudonymizing module allows the patients to be de-identified keeping just the demographic information related to three quasi-identifiers whose degree of specification is selected by the client from among the following options:

- Gender: a) removed, b) included
- Date of birth: a) removed, b) groups of 10 years, c) groups of 5 years, d) year, e) month, f) day
- Place of residence: a) removed, b) country, c) state, d) city, e) postal or zip code f) all included

The interactions among the clients, the pseudonymizing module and the demographic server are shown in points (a) through (f) in the text and in Fig. 4 and 5.

The pseudonymization of the extracts 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 to see whether it is already stored (b) in the associated demographic server (by means of its series of identifiers). In the affirmative case, it goes on to update, if necessary, the series of identifiers (c) that enter this entity into the demographic server with 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 identifiers of the entities of the extract. In the clinical part of the extract there are several fields linked by type *II* identifiers which refer to demographic entities which 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 which must equally be anonymized, such as the *party* field of the *RELATED_PARTY* class and the *performer* field of the *FUNCTIONAL_ROLE* class (see Figure 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 (*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 which allows access to the demographic information in free text fields. For this reason, a mechanism has been 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.

```
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. Pseudo-code of 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.

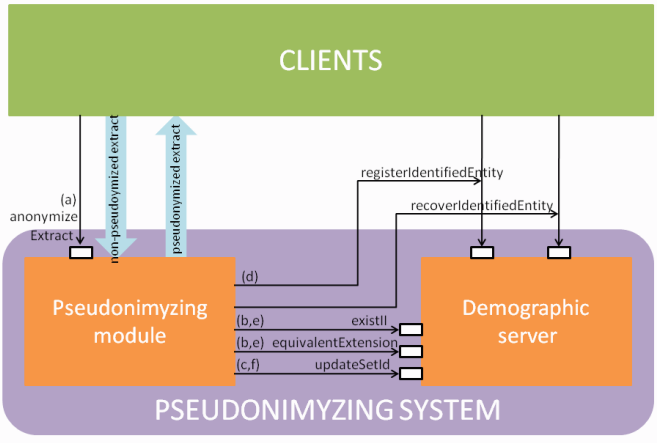

Fig. 5. Interaction between modules and clients in the pseudonymization process

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 has been found in the pseudonymized information.

In those cases in which the emitting system has also wanted to recover demographic information from the identifiers of the pseudonymized extract, it has been achieved successfully. In the above mentioned tests there has also been confirmed that the same identifier has been 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 article, 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 [30], 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 9,770 elements and the least numerous is that of men between 75 and 79 years' old with 1,100 integrants. Assuming a uniform distribution in these 5 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 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.

```
archetype (adl_version=1.4)
    CEN-EN13606-COMPOSITION.Other_demographic_data.v1

concept
    [at0000]

language
    original_language = <[ISO_639-1::en]>

definition
    COMPOSITION[at0000] occurrences matches {1..1} matches {  -- Other demographic data
        content existence matches {0..1} cardinality matches {0..1; unordered} matches {
            ENTRY[at0001] occurrences matches {0..*} matches {  -- Birthtime range
                items existence matches {0..1} cardinality matches {1..1; unordered} matches {
                    ELEMENT[at0002] occurrences matches {1..1} matches {  --
                        value existence matches {0..1} matches {
                            IVLTS[at0003] occurrences matches {1..1} matches {  -- IVLTS
                                low existence matches {0..1} matches {
                                    TS[at0004] occurrences matches {0..1} matches {  --
                                        time existence matches {1..1} matches {*}
                                    }
                                }
                                high existence matches {0..1} matches {
                                    TS[at0005] occurrences matches {0..1} matches {  --
                                        time existence matches {1..1} matches {*}
                                    }
                                }
                            }
                        }
                    }
                }
            }
        }
    }
```

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

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

Even though the current configuration of the pseudonymizer is the one described above, there exists another possible 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 non-anonymized data are being transmitted.

Other limitations of the system include: multimedia data (DICOM images, *attested_view* screenshots) in the extracts are not anonymized; apparitions of other identifiers (for instance pseudonyms) occurring in textual fields not included in the *demographic_extract* or not registered in the demographic server are also not detected. 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 owners of the clinical data sent. In this way, it facilitates compliance with current legislation on the protection of data 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. This 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 information of the demographic 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 work, 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 unit [31] together with other services such as: the storage and recovery of ISO/EN13606 extracts, the validation of extracts, archetype editor and server [32] 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 [33] for the extraction of clinical knowledge.

REFERENCES

[1] N. M. O´Boyle, R. Guha, E. L. Willighagen *et al*, "Open Data, Open Source and Open Standards in chemistry: The Blue Obelisk five years on" *J Cheminform*, vol. 3, no. 1, pp. *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, pp. *e308*. 2007.

[4] B. S. Elger, J. Iavindrasana, I. L. Lo *et al*, "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 *et al*, "Using EHRs to integrate research with patient care: promises and challenges" *J Am Med Inform Assoc*, vol. 19, no. 5, pp. *684-687*. Sep, 2012.

[6] International Organization for Standardization. ISO 13606 electronic health record communication part 1: reference model. ISO 13606-1.

[7] P. Libin, G. Beheydt, K. Deforche *et al*, "RegaDB: community-driven data management and analysis for infectious diseases" Bioinformatics, vol. 29, no. 11, pp. *1477-1480*. Jun, 2013.

[8] Spanish Law 14/2007 on Biomedical research. (accessed Jan 2014) [Online] Available: http://www.boe.es/boe/dias/2007/07/04/pdfs/A28826-28848.pdf

[9] Health Information Privacy. (accessed Jan 2014) [Online] Available: http://www.hhs.gov/ocr/privacy

[10] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule" *J Am Med Inform Assoc*, vol. 17, no. 2, pp. *169-177*. Mar, 2010.

[11] S. M. Meystre, F. J. Friedlin, B. R. South, *et al*, "Automatic de-identification of textual documents in the electronic health record: a review of recent research" *BMC Med Res Methodol*, vol. 10, pp. *70*. 2010.

[12] K. El Emam, L. Arbuckle, G. Koru *et al*, "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset" *J Med Internet Res*, vol. 14 no. 1, pp. *e33*. 2012.

[13] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity" *J Am Med Inform Assoc*, vol. 15, no. 5, pp. *627-637*. Sep, 2008.

[14] 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, pp. *e14*. 2012.

[15] L. Sweeney, "Simple Demographics Often Identify People Uniquely". Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000 (accessed Jan 2014) [Online] Available: http://dataprivacylab.org/projects/identifiability/paper1.pdf

[16] CEN. European Committee for Standardization. (accessed Jan 2014) [Online] Available: http://www.cen.eu

[17] International Organization for Standardization. ISO/IEC 10746-3. Information technology - Open Distributed Processing - Reference Model: Architecture.

[18] T. Beale, "Archetypes: constraints-based domain models for future-proof information systems". (accessed Jan 2014) [Online] Available: http://www.openehr.org/files/resources/publications/archetypes/archetypes_beale_oopsla_2002.pdf

[19] Health Level Seven International. (accessed Jan 2014) [Online] Available: http://www.hl7.org

[20] OpenEHR. (accessed Jan 2014) [Online] Available: http://www.openehr.org

[21] CIMI Wiki. Clinical Information Modeling Initiative. (accessed Jan 2014) [Online] Available: http://informatics.mayo.edu/CIMI

[22] A. Munoz, R. Somolinos, M. Pascual *et al*, "Proof-of-concept design and development of an EN13606-based electronic health care record service" *J Am Med Inform Assoc*, vol. 14, no. 1, pp. *118-129*. Jan, 2007.

[23] MySQL. [Online] Available: http://www.mysql.com

[24] Extensible Markup Language (XML). [Online] Available: http://www.w3.org/XML

[25] XML Schema. [Online] Available: http://www.w3.org/XML/Schema

[26] Java Persistence API. [Online] Available: http://en.wikipedia.org/wiki/Java_Persistence_API

[27] JAXB. [Online] Available: http://jaxb.java.net

[28] Web Services Activity. [Online] Available: http://www.w3.org/2002/ws

[29] Apache Tomcat. [Online] Available: http://tomcat.apache.org

[30] Population structure of Fuenlabrada. (accessed Jan 2014) [Online] Available: http://ayto-fuenlabrada.es/recursos/doc/SC/Estadisticas_y_territorio/36781_1111112013133426.pdf

[31] Interoperability services platform. (accessed Jan 2014) [Online] Available: https://hce13606.telemedicina.isciii.es:8443/interServer

[32] R. Sanchez-de-Madariaga, A. Munoz, J. Caceres *et al*, "ccML, a new mark-up language to improve ISO/EN 13606-based electronic health record extracts practical edition" *J Am Med Inform Assoc*, vol. 20, no. 2, pp. *298-304*. Mar, 2013.

[33] I. H. Witten, E. Frank and A. H. Mark, "Data Mining. Practical Machine Learning Tools and Techniques", *3*rd ed. Elsevier Ltd, Oxford; 2011