Ontology-Based Obfuscation and Anonymisation for Privacy

Leonardo H. Iwaya, Fausto Giunchiglia, Leonardo A. Martucci, Alethia Hume, Simone Fischer-Hübner, Ronald Chenu-Abente

Ontology-based Obfuscation and Anonymisation for Privacy

A Case Study on Healthcare

Leonardo H. Iwaya [1], Fausto Giunchiglia [2], Leonardo A. Martucci [1], Alethia Hume [2], Simone Fischer-Hübner [1], and Ronald Chenu-Abente [2] *

[1] Karlstad University, Sweden
[firstname.lastname]@kau.se
[2] Trento University, Italy
{hume,chenu,fausto}@disi.unitn.it

Abstract. Healthcare Information Systems typically fall into the group of systems in which the need of data sharing conflicts with the privacy. A myriad of these systems have to, however, constantly communicate among each other. One of the ways to address the dilemma between data sharing and privacy is to use data obfuscation by lowering data accuracy to guarantee patient's privacy while retaining its usefulness. Even though many obfuscation methods are able to handle numerical values, the obfuscation of non-numerical values (e.g., textual information) is not as trivial, yet extremely important to preserve data utility along the process. In this paper, we preliminary investigate how to exploit ontologies to create obfuscation mechanism for releasing personal and electronic health records (PHR and EHR) to selected audiences with different degrees of obfuscation. Data minimisation and access control should be supported to enforce different actors, e.g., doctors, nurses and managers, will get access to no more information than needed for their tasks. Besides that, ontology-based obfuscation can also be used for the particular case of data anonymisation. In such case, the obfuscation has to comply with a specific criteria to provide anonymity, so that the data set could be safely released. This research contributes to: state the problems in the area; review related privacy and data protection legal requirements; discuss ontology-based obfuscation and anonymisation methods; and define relevant healthcare use cases. As a result, we present the early concept of our Ontology-based Data Sharing Service (O-DSS) that enforces patient's privacy by means of obfuscation and anonymisation functions.

1 Introduction

In today's Information Society, people are surrounded by information technology in their everyday life. Providers of information services often record and categorize people, or data subjects, into *profiles*. Profiles consist of personal data

* This research was funded by SmartSociety, a research project of the Seventh Framework Programme for Research of the European Community under grant agreement no. 600854.

that is managed, shared and modified by different information systems; often without the individual's consent [4]. To protect the subject's rights over their personal data, security and privacy are imperative in the design of solutions that handle sensitive information. Security is commonly addressed by means of the principles of confidentiality, integrity, and availability. Privacy, in turn, stands for fundamental rights and freedoms of subjects to have their right to privacy with regards to the manipulation and processing of personal data [1].

Among current technologies, Healthcare Information Systems (HIS) are frequent target of information security and privacy researches. The reasons are manifold. HIS are essential and widely-deployed systems that manage highly sensitive data; providers have to comply with security/privacy regulations; and, data breaches might cause expensive penalties and damage to the company's reputation. Notwithstanding, the patient's records have to be shared among multiple healthcare service providers, either for primary and secondary purposes. For instance, Electronic Health Records (EHR) might be distributed, within affiliated hospitals and medical centers (i.e., inter-institutional EHR). In this case, medical data is exchanged for primary use, i.e., *meaningful use* for patient's treatment, with an implied trusted domain and confidentiality among medical staff. However, EHR are also increasingly being used for secondary purposes, such as release of data for governmental health programs and research [5]. EHR can also be integrated to Personal Health Records (PHR)(e.g., HealthVault[3] and PatientsLikeMe[4]), and consecutively linked to all sorts of patient-centered and patient-controlled information systems (e.g., mobile healthcare). These multiple data flows add further concern regarding security and privacy.

One way, that we focus in this paper, to cope with the dilemma between data sharing and data privacy refers to the use of abstractions; in particular the use of obfuscation and anonymisation. In our research we are considering the concept of abstraction as a broader field – that remains to be understood. By abstraction we mean, the process of mapping a problem representation onto a new one, preserving certain desirable properties and reducing its complexity [8]. Particular cases of abstraction studied here are obfuscation and anonymisation.

By obfuscation we mean, to lower individual data item accuracy in a systematic, controlled, and statistically rigorous way [2]. By anonymisation, we intent to protect privacy by making a number of data transformations so that individuals whom the data describe remain anonymous [5]. In this case, data transformations are essentially obfuscation functions that can achieve anonymity. Anonymity, in turn, is a property of an individual that cannot be identified within a set of individuals, the *anonymity set* [15]. The anonymisation process can have variable degrees of robustness [19], depending on how likely is to: 1) single out an individual in the dataset; 2) link records concerning the same individual; or, 3) infer the value of one attribute based on other values. Therefore, we claim that

[3] Microsoft HealthVault (www.healthvault.com)
[4] PatientsLikeMe (www.patientslikeme.com)
[5] Anonymous: someone unknown; not distinct; or, lacking individual characteristics.

anonymisation is a special case of obfuscation; and accordingly, obfuscation is a special case of abstraction.

In this paper, we aim to investigate the problem of data obfuscation and its particular case of anonymisation, through the use of a privacy-enhancing ontology-based obfuscation mechanism for releasing PHR and EHR to selected audiences with different degrees of obfuscation. Data minimisation and access control are supported, as different actors, e.g., doctors, nurses and managers, will need to get access to the just amount of information needed for their tasks. In addition, an ontology-based obfuscation can be used to decrease the semantic loss, i.e., maintain a high degree of utility of the anonymised data. This research contributes to: stating the problems in the area; reviewing privacy and data protection legal requirements; discussing ontology-based obfuscation and anonymisation methods; and defining relevant healthcare use cases. As a result, we present the early concept of our Ontology-based Data Sharing Service (O-DSS) that enforces patient's privacy by means of obfuscation and anonymisation functions.

This paper is organized as follows. In Section 2 we briefly motivate this research with respect to the legal aspects of privacy and data protection regulations and legislations around HIS. In Section 3 we provide a summary of the relevant terminology, existing methods for data obfuscation and anonymisation, and related work on ontology-based approaches. In Section 4 we introduce a preliminary design of our privacy-preserving O-DSS. In Section 5 we discuss the research future work, and in Section 6 we present our conclusions.

2 Data Protection Regulations and Legislation

The European legal privacy framework that we will in this section refer to is based on the EU Data Protection Directive 95/46/EC [1] and the upcoming General EU Data Protection Regulation (GDPR) [6], which was approved by the European Council in June 2015 and is expected to replace the national laws implementing the Directive in the near future (probably in 2016).

For the Health Care section, there is no specific harmonized EU data protection legislation, as this is rather regulated by different national legislation that takes consideration of the national different heath care practices.

Still, the general rules of the Directive and soon of the GDPR will apply unless there are overriding (national or EU) legal rules.

Of particular importance for motivating our work is the general privacy principle of data minimisation included in the Data Protection Directive and Regulation, for instance cf. Art. 5 (c), (e) GDPR: Personal data should be: *"(c) adequate, relevant, and limited to the minimum necessary in relation to the purposes for which they are processed"*; *"they shall only be processed if, and as long as, the purposes could not be fulfilled by processing information that does not involve personal data (data minimisation)"*; and *"(e) kept in a form which permits direct or indirect identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed"*;

Furthermore, the Art 23 of the GDPR requires that the data controller should follow a *Data Protection by Design and by Default* approach by implementing *"technical and organisational measures appropriate to the processing activity being carried out and its objectives, such as data minimisation and pseudonymisation..."*.

Besides the European framework, in the United States the Health Insurance Portability and Accountability Act of 1996 (HIPAA), is also another well-known example of regulatory mechanism for privacy. Some of the most traditional methods (e.g., k-anonymity) as well as the HIPAA's standards for anonymisation/de-identification are discussed in Section 3.

3 Background and related work

According to [8], the process of abstraction relates to the process of separating, extracting from a representation another "abstract" representation, which consists of a *brief sketch* of the original representation. Therefore, the same authors were able to informally define abstraction as: *the process of mapping a representation of a problem onto a new representation, which helps to deal with the problem in the original search space by preserving certain desirable properties, and, is simpler to handle.* This concept was originally define in the field of artificial intelligence, in which abstraction refers to reasoning. Likewise, the objective of obfuscation also incorporates the very same elements. Obfuscation is a reasonable transformation of the data that preserves certain properties (e.g., semantics, analytics, statistics), and typically entails generalization. That is why we also define obfuscation (and anonymisation) as special cases of this broader theory of abstraction.

For this research, we are particularly interested in abstractions that can be supported by ontology-based knowledge representations to either obfuscate or anonymise values. In addition, we also consider practical aspects of the healthcare field, such as: clinical vocabularies and ontologies that are employed in the data structures used in EHR. This and other concepts that are grounding our work are explained and briefly discussed in this section.

3.1 EHR's data elements

Medical standards for EHR vary from one country to another, but the Health Level Seven (HL7) and Comite Europeen de Normalization – Technical Committee (CEN TC) 215 are probably the most renowned ones. The HL7 group develops the most widely messaging standard for healthcare in United States. The CEN CT 215 operates in 19 European member states and is the main healthcare IT standards in Europe. The work of such organizations is important to achieve interoperability among HIS, which for EHR refers to the definition of clinical vocabularies, message exchange formats, and EHR ontologies. In the present research, however, we are only interested in how to exploit the structured vocabularies and ontologies during the anonymisation process.

In brief, clinical vocabularies or standard vocabularies are used to agree upon the use of medical terminologies when writing in a patient record. All the terms are usually encoded in order to facilitate data exchange, comparison, or aggregation among HIS. Some established examples are the International Classification of Disease (ICD[6]) and the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT). Nevertheless, data elements inside EHR may also be in free-form text, for example, clinician's notes. Any anonymisation method for EHR needs therefore to cope with structured data elements as well as form-free running text.

3.2 Conventional anonymisation methods

The k-anonymity, formulated by Sweeney [17], was one of the first and well-known formal methods that address the issue of data anonymisation. In a more formal definition, the initial scenario consists of a data holder that held a collection of person-specific, field structured data; and wants to share a version of this data with researchers. The data holder considers the following problem [17]: "*How can the data holder release a version of its data with scientific guarantees that the individuals who are the subjects of the private data cannot be re-identified while data remain practically useful?*" The solution is the k-anonymity property, which means that the information for each person contained in the released data cannot be distinguished from at least $k-1$ individuals whose information also appear in the released data. In brief, anonymised data with k-anonymity property guarantee an anonymity set size of $k-1$.

To do so, the program receives as input a table with n rows and m columns, in which each row of the table represents a person-specific record of the population and the entries in various rows should not be unique. The algorithm than applies two different methods to achieve k-anonymity:

- Suppression, certain values have to be simply replaced by an asterisk $'*'$. For instance, person direct *identifiers* should be omitted (e.g., name, address, phone, personal numbers).
- Generalization, individual values of attributes are replaced by broader category. For instance, an attribute 'age' can be replaced by a range of values, i.e, age $'21'$ by $' \geq 20'$.

The k-anonymity offers a straightforward and to some extent effective method, but the approach is still susceptible to homogeneity and background knowledge attacks, leading researchers to the design of improved version, such as l-diversity [11] and t-closeness [10]. All these methods are, however, somewhat naive when dealing with non-numerical values and are unable to maintain enough semantic coherence after anonymisation [12]. To cope with semantics, many authors proposed different ontology-based anonymisation methods, further explained in the next section.

[6] http://www.who.int/classifications/icd/en/

Besides the formal methods for data anonymisation, we could also consider heuristic-based strategies. For instance, with respect to the HIPAA Privacy Rule [14], the EHR can be de-identified using the "*Safe Harbor*" standard. In this case, a number of 18 types of identifiers that compose the Protected Health Information (PHI) should be removed. Safe Harbor is a very simple approach, but it does not provide any scientific guarantees for anonymity sets nor protection to re-identification attacks [3]. Fortunately, the HIPAA Privacy Rule also considers a second standard called "*Expert Determination*", which means the application of statistical or scientific principles to reduce re-identification to a very small risk.

3.3 Ontology-based approaches

An ontology is a method for knowledge representation, which uses a formal, explicit and machine readable structure of concepts hierarchically interconnected by a semantic network [7]. These powerful data structures enable knowledge organization, sharing, and emulation of cognitive processes and / or common understandings of specific domains. We are particularly interested in the concept of semantic obfuscation of ontology-based systems that can be used to restrict the release of information according to the audience. A few ontology-based privacy-preserving mechanisms were recently proposed. In this section, we made an effort to summarize and briefly evaluate the findings of these studies.

Access control and context obfuscation. In a more generic approach, the work of [18] proposes a *context obfuscation* mechanism for pervasive networking and context-aware programs. The system allows the users to set different privacy preferences and stipulate rules to control the access of context information. In brief, all the attributes related to an user can have different access control settings for information, depending on the context, requesters, and use purposes. Besides the privacy preferences, the user can also define granularity levels of access, which is based on ontological structures that can capture the granularity relationship between instances of an object type.

For example, when the patient Alice wants to give to her doctor Bob access to her attribute Diagnosis, she configures her privacy preferences as follows:

```
privacy_pref_list.add_rule(
   consumer = Bob,
   attribute = Diagnosis,
   purpose = Treatment,
   allow = True )
```

In addition, Alice can set the generalization level $l \in \mathbb{Z}$ that should be applied to the attribute. The higher the level, the higher the generalization level would be, i.e., going upwards in the ontology (for instance, see Figure 5).

```
generalization_pref_list.add_rule(
   consumer = Bob,
```

```
attribute = Diagnosis,
level = 0 )
```

This context obfuscation mechanism was already used in [16], to provide a privacy-preserving and granular access control to PHR. The authors, however, are still missing the link with real ontologies and medical vocabularies, which could greatly improve the obfuscation quality in real HIS systems.

Ontology-Based Anonymisation. An ontology-based data set anonymisation with categorical (i.e., textual) values is proposed by Martínez et al. [12,13]. They aim at preserving data semantics of anonymised values. The proposal relies on a set of heuristics to optimize obfuscation, and ensure scalability in cases of heterogeneous data sets and wide ontologies. In addition, their algorithm also employs k-anonymity to provide a minimum set of privacy guarantees. To do so, the solution relies of the measurement of the semantic similarity, i.e., to quantify the taxonomical resemblance of compared terms based on a knowledge base. Therefore, it is possible to semantically compare, rank and group the most similar record values. Subsequently, the method aggregates values in a group, which refers to the process of replacing the values in several records by a single one, summarizing and making them indistinguishable (i.e., k-anonymity set). This operation performed by means of the author's proposal of a centroid calculus for multi-variate non-numerical data, to obtain accurate centroids in a group (for further details we refer the reader to [12,13]). Their use case [13] provides a more concrete example of EHR anonymisation, including many categorical values from SNOMED CT, which makes the method's applicability more realistic.

3.4 Ontology-based Identity Management and Access Control

In [9], the authors introduce the concept of a privacy-enhanced Peer Manager, in which the original idea was designed to preserve privacy in collective adaptive systems. The Peer Manager works as an user-centered identity management platform that keeps user's information private. This framework was built upon the privacy policy language PPL (PrimeLife Policy Language), with which every user can control his personal information by imposing access and usage control restrictions. As a privacy-enhancing structure, this platform instead of directly allowing access to peer's information, creates a *Profile* structure that are sent as replies to queries. The created Profile, in turn, reveals only partial or obfuscated information about the Entities. In essence, the Peer Manager never discloses the Entities' original data, but derived Profiles previously defined by the users. Besides, if compared to [18], the Peer Manager supports a far more general and robust approach for access control and obfuscation.

3.5 Putting things together

The approaches discussed here employ obfuscation in different scenarios yet with similar goals. In summary, we aim to integrate the proposals [13] and [18] into

the Peer Manager [9] obfuscation functions, and thus, demonstrate how it can be applied to healthcare systems (e.g., EHR and PHR) using real medical ontologies.

4 Obfuscation and anonymisation for HIS

Health information is, in general, managed by systems for primary purposes, i.e., the provisioning of health care to the benefit of the patient. It is noteworthy, however, that aforementioned requirements are still valid, such as trusted medical environment, with implied confidentiality among healthcare workers. The medical institutions usually are the data custodians in case of EHR, and thus, any data breach is the institution's responsability. Security mechanisms in the EHR should provide confidentiality, integrity, availability, and make personnel accountable for unauthorized data release, by means of logging and auditing tools.

Nevertheless, health care services also use the health information for secondary purposes, such as general public health monitoring, evaluation of healthcare programs, and research. In the case of a secondary use, data should be subject to *de-identification* or *anonymisation*, such as the aforementioned Safe Harbor and Expert Determination standards from HIPAA. In this section we explain how obfuscation and anonymisation can be used in the healthcare context.

4.1 Ontology-based Data Sharing Service

We propose an ontology-based data sharing service (O–DSS) to mediate access to healthcare data sets. Many HIS applications fit in this scenario. In this preliminary research we focus on Semantic Obfuscation (SO) and Data Anonymisation (DA) for standard HIS, such as EHR and PHR. In brief, we consider the following use cases and their information flows:

1. Primary Use
   - EHR $\rightarrow$ O–DSS (SO) $\rightarrow$ privacy-preserving patient treatment (hospital and clinics).
   - PHR $\rightarrow$ O–DSS (SO) $\rightarrow$ patient's granular control of own health.
   - EHR $\cup$ PHR $\rightarrow$ O–DSS (SO) $\rightarrow$ reminder or alert systems for family or caregivers.
2. Secondary Use
   - EHR $\cup$ PHR $\rightarrow$ O–DSS (SO + DA) $\rightarrow$ medical research repository.
   - EHR $\cup$ PHR $\rightarrow$ O–DSS (SO + DA) $\rightarrow$ nationwide HIS network.

Each information flow refers to a different branch of this use case (examples of flows 1, 2, and 4 are depicted in Fig. 1). Therefore, the SO and DA techniques should deal with different requirements, according to the target application of the data. In what follows, we provide further details of each use case.

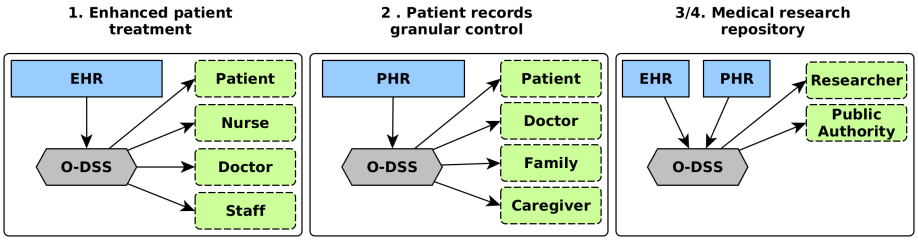


Fig. 1: A healthcare use case for ontology-based obfuscation for HIS. The dashed boxes are the data requesters $r \in R$, and EHR and PHR are the data providers $p \in P$. The ontology-based data sharing service obfuscates data from p that is communicated to r.

UC1: Privacy-preserving patient treatment Patient's data can be accessed by clinicians, nurses, secretaries, and accountants for many purposes inside the hospital. All the employees are making meaningful use of the data, and therefore, they are under an implied confidentiality agreement. Furthermore, medical institutions also are allowed to share EHR among affiliated institutions and healthcare services – still, the institution is liable for the data's confidentiality. In such cases, it is more important to enforce access control and, in a privacy-preserving perspective, apply data minimisation whenever possible to reduce risks of data leakage.

UC2: Patient's granular control of E/PHR PHR are being increasingly used by patients to track their own daily activities (e.g., wellness and fitness applications), or to have an interface to their EHR (e.g., patient web portals or dashboards). In particular, if patients transfer their data to private services (non-medical) that support the management of health records (e.g., Microsoft HealthVault), than, the medical institutions might not be liable for the secrecy of released data. This user-centered applications encourage patients to have more control of their own health, and also, provide means to granularly share PHR with other healthcare services, social networks, family, and so on. In this case, we consider that patient consents with the data release, but mechanisms should provide the patient with granular control in the form of selective disclosure and obfuscation options in dependence on the different data consumers.

UC3: Reminder or alert systems Reminder systems are commonly used for drug treatment compliance, and can be linked to EHR, to provide reminders to out-patients or chronic patients. Some alert mechanisms also exist, providing EHR access to family members and caregivers in case of emergencies. As presented in Figure 1, the data flows might come from EHR or PHR, since there are private non-medical services that support this applications. This use case has privacy requirements that are similar to UC2, since the user can have directly configure access control settings.

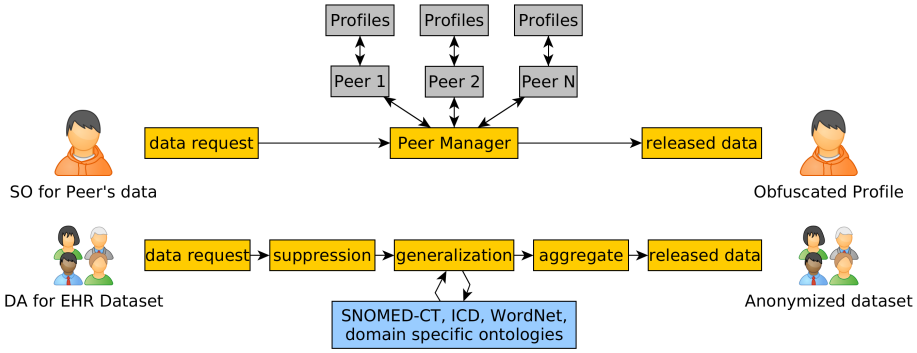


Fig. 2: O–DSS processes for SO and DA.

UC4: Medical research repository EHR are frequently used as source of clinical information for medical-related research. To do so, EHR are usually de-identified and/or anonymised before releasing the data, in which cases there would be no need of patient's consent; exceptional cases of non-medical research, e.g., marketing or financial studies.

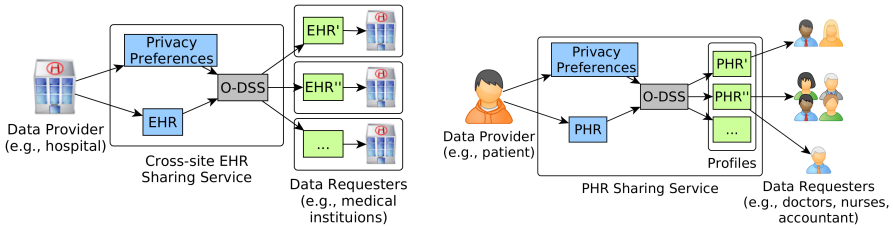
UC5: Nationwide HIS network Similarly to UC4, more ambitious projects aim to create nationwide HIS networks, that would interconnect EHR systems within a country (i.e., primary use), or even, medical repositories for research (i.e., secondary use) – also knows as translational research information system (TRIS). For instance, in [5], the authors examine the privacy issues on building a database integrating clinical information from an EHR systems with a DNA repository.

4.2 SO and DA functions

The O–DSS provides two fundamental functions: semantic obfuscation (SO) and data anonymisation (DA). The SO function is specially grounded on the proposals: peer profiling [9] and context obfuscation [18]. That is, we aim to partially show or obfuscate the record (i.e., to provide data minimisation instead of anonymisation), based on the concept of Peer Manager for access control, and also, exploit the medical ontologies for data obfuscation.

The DA is grounded on k-anonymity [17] and improved techniques presented in [12,13]. Figure 2 illustrates how the O–DSS manages the data flows to provide SO and DA. Besides, in Figures 3 and 4, O–DSS is placed into context in line with aforementioned use cases.

Primary use and semantic obfuscation. In the case of primary use, the semantic obfuscation (SO) acts as a data minimization mechanism. The objective is to restrict the amount of health information that should be disclosed for a



(a) (b)

Fig. 3: Data release/sharing for (a) cross-institutional patient treatment and (b) user-centered (ontology-based) obfuscation and granular access control for PHR.

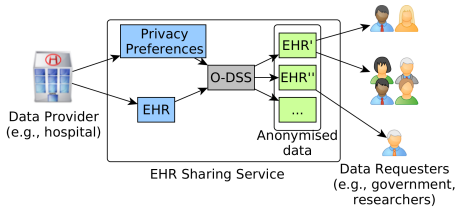


Fig. 4: Data release/sharing to create research repositories.

specific task (e.g., visualize a patient record or check a doctor's schedule) and to a specific person or role (e.g., a nurse) to the required minimum. Healthcare Personnel from different Departments will have different access right: nurses will only have access to EHR of patients that they are treating; psychiatric diagnoses, might only be seen by psychiatrists, but not by other treating doctors (who may however see that a psychiatric diagnosis exists); and, values about blood infection/HIV would be read by all Healthcare Personnel for employee's security reasons.

We define two sets of actors: data *providers* (P), i.e. the data subject (patient, PHR user) and data *requesters* (R), where R makes queries to P about specific health information.

Alice $\in P$ is first registering with the Peer Manager that acts as a data controller and enforces the *Alice*'s privacy preferences on her behalf. The Peer Manager provides *Alice* with a set of PPL policies for different peer profiles representing partial identities that the Peer Manager should manage on *Alice*'s behalf. *Alice* can either choose from this set of policies or construct her own policies for profile. For enabling that *Alice* can determine different semantic obfuscations for different "audiences", PPL is extended with obligations to apply different obfuscation operations at different granularity along a specified ontology hierarchy in dependence of different data requesters or roles of data requester (that correspond to different "downstream controllers" in PPL terminology). *Alice* is sending his profiles together with the PPL policies that should apply for these profiles to the Peer Manager.

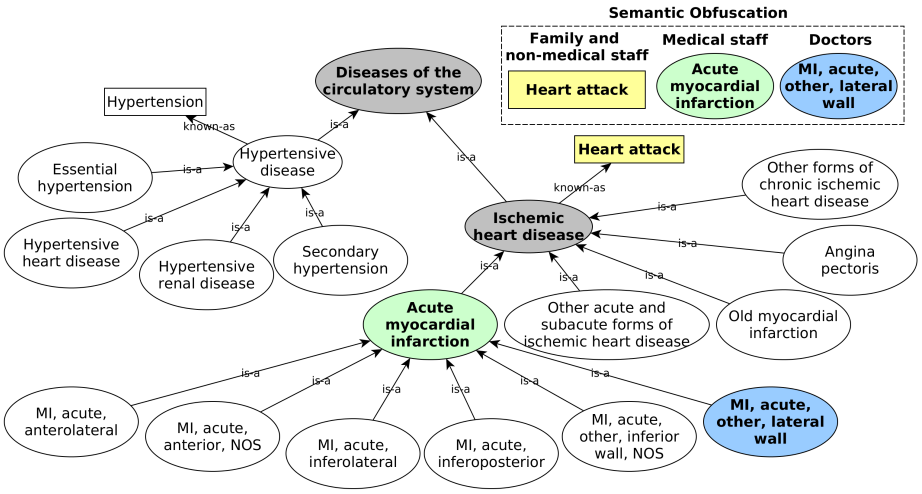


Fig. 5: Example of ontology-based obfuscation.

Once when $Bob \in R$ makes queries to *Alice* about specific health information, following the approach in [9], the request has to be mapped into a profile of $Alice \in P$ that will be returned to *Bob*. Hence, before replying, the Peer Manager would check the access rights that *Alice* has given to *Bob*, i.e., the profile that *Alice* has decided to reveal to *Bob* for a given task and purpose (according to the PPL policy that *Alice* has defined or chosen for that profile). Furthermore, if the access conditions are fulfilled, i.e. data is to be forwarded to a so-called downstream controller (*Bob*), the obfuscation obligations that were defined in the policy for the event of data forwarding are first triggered by the Peer Manager: Within the revealed profile (i.e., an attribute-based description of *Alice*), the ontology-based semantic obfuscation is applied to each attribute, allowing *Alice* to obfuscate its data with different granularity levels (i.e., different semantic level) according to the obfuscation obligations defined in the profile's PPL policy for downstream controller *Bob* (or for his role). Thus, as shown in Figure 6, the Peer Manager follows the PPL obligations, e.g., executing the obfuscation functions accordingly.

Different attributes with different types of values will require appropriate ontologies/mechanisms to obfuscate them. The Figure 5 shows an example of how the ontology-based obfuscation can be used on an attribute describing diagnosis of a patient in order to abstract the information revealed to different requesters (namely, family member, medical staff, or doctor). Another way to perform semantic obfuscation in a patient's profile is by revealing different subsets of attributes (i.e., complete obfuscation of some attributes). For instance, a doctor would be able to retrieve the list of prescribed medications from a patient while a hospital accountant would only see the aggregate drug cost. Hence, *Bob*'s

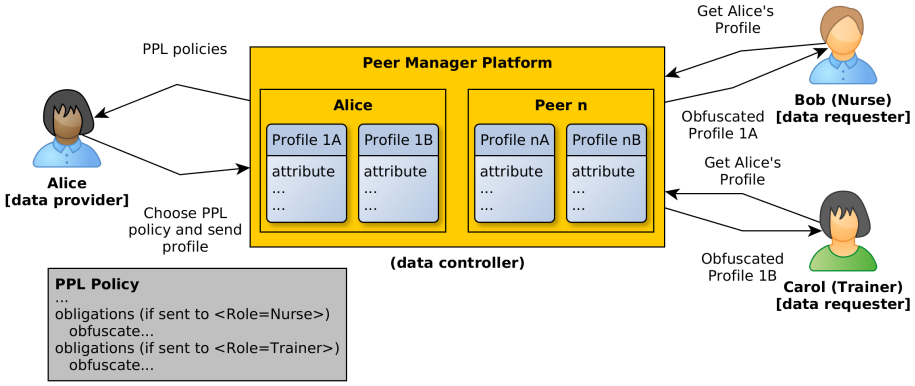


Fig. 6: Obfuscation process with Peer Manager and PPL.

access rights determine the accuracy of how data is accessible, i.e. the level of achieved obfuscation.

Some important distinctions from the work of [18], we did not emphasize the use of *context*, which would allow *Alice* to further refine her privacy preferences based on her location and current activity. Nonetheless, our proposal considers that the SO function should exploit real medical vocabularies, such as ICD-10 and SNOMED-CT, instead of using *ad-hoc* or domain specific ontologies. Moreover, we integrate our mechanism with the Peer Manager [9], which provides a more sophisticated access control model.

Secondary use and data anonymisation. If an entire data set of health information has to be shared for a secondary purpose with a data processor, such as a public health organization, an ontology-based semantic obfuscation / anonymisation can maximize data utility by preserving the data semantics while eliminating personal information from the data set to the degree required, i.e., the mandatory data anonymisation (DA) process. The data consumer is usually a third party, such as a research institution or a public service that need the data for secondary purpose (see Figure 4).

Also in this case the general notion of abstracting the information of profiles, as proposed in [9], can be applied. The main difference is that a data consumer is now asking for the whole set of, for instance, patients profiles. Before replying, the peer manager should be able to find/compute a profile for each peer (i.e., patient) in the data set such that certain level of anonymity is guaranteed. For example, if the peer agreed to reveal information for a secondary purpose provided that $k-$anonymity is guaranteed, then the anonymisation process has to be applied (i.e., follow PPL obligations) to attributes of profiles until such requirement is achieved. Considering again the example from Figure 5 and the attribute describing the diagnosis on each peer's profile, the anonymisation process needs to select the level of detail to be included in the revealed profiles such that a

given patient's record can not be re-identified by their diagnosis. For these cases, the ontology is important to improve the data utility during the mandatory data anonymisation (DA) process. Here we adopt the solution presented in [13], that enables $k-$anonymisation of structured non-numerical medical retaining semantics by using SNOMED-CT as knowledge base.

Moreover, by exploiting real medical vocabularies in the ontology-based obfuscation the approach becomes more robust and usable in real scenarios dealing with E/PHR. Another important feature to highlight is that a solution designed in this way is scalable in terms of the underlying ontology being used, i.e., the ontology can change, evolve or grow while the above approach is still applicable.

5 Future Work

Currently, we have mainly positioned how ontologies-based obfuscation and anonymisation can be used in HIS; by addressing legal requirements, reviewing many of the existing methods and putting them into the context of HIS. We also show how SO and DA can be used together with the Peer Manager and PPL. Notwithstanding, we noticed that concepts could be refined, and the link between theory of abstraction and obfuscation can be further formalized. In a broader sense, we aim to understand how the areas the privacy and the ontology areas could cooperate, in order to support data privacy.

Apart from that, future work has many challenges that remain to be addressed. Currently, we are not discussing the usability issues for setting all the privacy preferences for SO and DA. In addition, regarding obfuscation, the problem of inferences by correlations (e.g., infer the patient's original disease given the list of drugs/medical procedures) is still open. And for DA, we are considering mainly k-anonymity (i.e., anonymisation by generalization), but other methods based on randomization (e.g., noise addition, differential privacy) are also worthy considering. A complete solution would make use of many different DA methods.

6 Conclusions

This paper presented the use of a privacy-preserving ontology-based obfuscation mechanism intended to obfuscate health information either for primary or secondary use. In the case of primary use, minimization of personal data means that an actor gets no more information than needed and with an appropriate semantic level. For secondary use, the proposed mechanism can minimize the semantic loss of data, such that a high degree of utility is maintained, while data is anonymised to the specified DA requirements. Additionally, we described five use cases to illustrate the O–DSS, and we discussed how it can be integrated with the existing Peer Manager. Obfuscation capabilities were expected in the PrimeLife and Smart Society (i.e., Peer Manager) projects, but there were no

clear examples on how to use them. The obfuscation functions can be implemented by extending the obligations in the PPL policy, defined between data requester and data controller, and thus, allowing SO and DA over the attributes.

Acknowledgments

The authors gratefully acknowledge Hans Hedbom for his assistance with PrimeLife Policy Language and reviews that helped to improve the manuscript. Furthermore, the authors also thank Rose-Mharie Åhlfeldt, anonymous reviewers and participants of IFIP Summer School (2015), whose comments and suggestions greatly contribute to enhance and clarify our work.

References

1. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L No.281 (23 Nov 1995)
2. Bakken, D.E., Parameswaran, R., Blough, D.M., Franz, A.A., Palmer, T.J.: Data obfuscation: Anonymity and desensitization of usable data sets. IEEE Security & Privacy (6), 34–41 (2004)
3. Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the hipaa privacy rule. Journal of the American Medical Informatics Association 17(2), 169–177 (2010)
4. Camenisch, J., Sommer, D., Fischer-Hübner, S., Hansen, M., Krasemann, H., Lacoste, G., Leenes, R., Tseng, J., et al.: Privacy and identity management for everyone. In: Proc. of the 2005 Workshop on Digital Identity Management. pp. 20–27. ACM (2005)
5. El Emam, K.: Methods for the de-identification of electronic health records for genomic research. Genome Medicine 3(4), 25 (2011), http://genomemedicine.com/content/3/4/25
6. EU Commission: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (2015), http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf
7. Fensel, D.: Ontologies. Springer (2001)
8. Giunchiglia, F., Walsh, T.: A theory of abstraction. Artificial Intelligence 57(2), 323–389 (1992)
9. Hartswood, M., Jirotka, M., Chenu-Abente, R., Hume, A., Giunchiglia, F., Martucci, L.A., Fischer-Hübner, S.: Privacy for peer profiling in collective adaptive systems. In: Camenisch, J., Fischer-Hübner, S., Hansen, M. (eds.) Privacy and Identity Management for the Future Internet in the Age of Globalisation, IFIP Advances in Information and Communication Technology, vol. 457, pp. 237–252. Springer International Publishing (2015), http://dx.doi.org/10.1007/978-3-319-18621-4_16
10. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 106–115 (April 2007)

11. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1(1) (Mar 2007), http://doi.acm.org/10.1145/1217299.1217302
12. Martínez, S., Sánchez, D., Valls, A.: Ontology-based anonymization of categorical values. In: Proc. of the 7^{th} Int. Conf. on Modeling Decisions for Artificial Intelligence (MDAI). LNCS, vol. 6408, pp. 243–254. Springer (Oct 2010)
13. Martínez, S., Sánchez, D., Valls, A.: A semantic framework to protect the privacy of electronic health records with non-numerical attributes. J. of Biomedical Informatics 46(2), 294–303 (2013)
14. OCR: Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Office for Civil Rights (2012), http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
15. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf (Aug 2010), http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf, v0.34
16. Rahman, F., Addo, I.D., Ahamed, S.I.: Prisn: A privacy protection framework for healthcare social networking sites. In: Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems. pp. 66–71. RACS '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2663761.2664199
17. Sweeney, L.: k-Anonymity: a Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
18. Wishart, R., Henricksen, K., Indulska, J.: Context obfuscation for privacy via ontological descriptions. In: Location-and Context-Awareness, pp. 276–288. Springer (2005)
19. WP29: Opinion 05/2014 on anonymisation techniques. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (Apr 2014)