MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190414

Layered Privacy Language Pseudonymization Extension for Health Care

Armin Gerl[a], Felix Bölz[a]

[a] *Chair of Distributed Information Systems, University of Passau, Passau, Bavaria, Germany*

Abstract

Enforcement of General Data Protection Regulation strengthens privacy in Europe and especially emphasizes protection of special categories of data as required in health care. Layered Privacy Language intends to model privacy policies to enforce them. Hereby, a special focus lays on the Policy-based De-identification process, which is based on anonymization and privacy models. Motivated by a health care scenario, this work shows pseudonymization capabilities are essential for health care. An overview of pseudonymization methods is given, showing a great variety of methods for different use cases. Therefore, a pseudonymization extension for Layered Privacy Language is introduced to define several pseudonymization methods. Furthermore, pseudonymization is added to Policy-based De-identification process of the overarching privacy framework of Layered Privacy Language. An example policy configuration is given demonstrating the introduced pseudonymization extension on the given health care example. The results are discussed, concluded, and future work is introduced.

Keywords:

Anonyms and pseudonyms, confidentiality, privacy

Introduction

The enforcement of the General Data Protection Regulation (GDPR) in May 25th 2018 strengthens privacy in Europe. The legal principles of Privacy by Design and Privacy by Default postulate that privacy has to be taken into consideration during design of all technical systems and default settings (e.g. privacy policy in an electronic system) have to be privacy friendly by default [Art. 25, 1]. Special categories of personal data, e.g. data concerning health or genetic data, are hereby especially protected and must not be processed unless special conditions are met like a legal basis or explicit consent [Art. 9, 1]. The Layered Privacy Language (LPL), including its overarching framework, intends to model and enforce privacy policies 'from consent to processing' [2]. Hereby, LPL models privacy policies, which are presented and possibly personalized by the individual Data Subject. If agreement and/or consent is given on the privacy policy, it will be pre-processed and stored. On a request of personal data for a specific purpose the Policy-based De-identification process is conducted. This determines if the requesting entity is authorized to request the data and, if necessary, anonymizes the data. Thus, anonymization techniques like generalization, suppression, and deletion are used to enforce privacy models like k-anonymity [3], l-diversity [4], or diffential privacy [5].

In this work, we introduce a scenario that demonstrates that privacy models solely are not suitable for health care. In health care, capabilities for pseudonymization have to be introduced in LPL, section *'Data Analysis in Health Care'*. Therefore, we give an overview of pseudonymization techniques in section *'Pseudonymization'*. In section *'Integration in LPL'* we describe extension of LPL, as well as integration of the pseudonymization techniques within existing Policy-based De-identification process. Section *'Health Care Scenario Results'* discusses extension of LPL according to the introduced scenario. In *'Discussion'* possibilities and limitations of the approach are shown. Lastly, the work is summarized and future work given in *'Conclusion'*.

Data Analysis in Health Care

In health care, patients' personal data is stored and processed for different purposes like billing, analysis, or research. Especially, information on patients' condition, e.g. blood sugar, symptoms, diseases and treatments, which fall under the special categories of personal data [Art. 9, 1], are stored alongside regular personal data like name, age or address, of patients. In the following, we assume a scenario in which a dataset is analysed to identify a possible epidemic. Here, influenza, commonly known as 'the flu', is divided into different types. *Type A* is generally responsible for large flu epidemics, *Type B* is less harmful, and *Type C* does not cause epidemics. There are additional sub-types, which are irrelevant for this example.

Table 1– Example health care dataset describing influenza types for patients. Attributes are assigned to privacy categories.

Name (*EI*)	**Age (*QI*)**	**Zip Code (*QI*)**	**Virus (*SD*)**
John	28	94032	Flu Type C
Max	25	94032	Flu Type C
Mary	22	94034	Flu Type A
Harry	20	94032	Flu Type B
Theresa	24	94034	Flu Type C

We assume a health care dataset like shown in Table 1, consisting of attributes of different privacy categories. In general, data is classified in four categories. *Explicit Identifiers (EI)*, attributes which identify a person uniquely. *Quasi Identifiers (QI)* which in combination identify a person. *Sensitive Data (SD)*, consisting of sensitive but not identifiable information about the record owner. *Non-Sensitive Data (NSD)*, data not in any of the categories above [6]. Based on this classification, which is also present in LPL, privacy models anonymize a dataset. We assume in this case that we want to protect the identity and sensitive information, which could also be maliciously utilized for de-identification [4]. Therefore, a

privacy model like l-diversity could be choosen and applied on the dataset.

The *EI* in l-diversity is anonymized by deletion (denoted by '*'). The *QI* and *SD* are anonymized, such that each group (age, zip code, and symptom) contains at least three identical entries. The values for attribute 'age' are generalized by decades, which leads to generalization of each value (e.g. '20' to '20 – 29'). For zip code, suppression is applied beginning with the last character, which is suitable for German zip codes. Lastly, attribute 'virus' is generalized from specific to more general description, e.g. 'Flu Type A' to 'Flu'.

The resulting anonymized dataset (see Table 2) may be interpreted as an epidemic in the general area of 9403* despite original data shows only one case of influenza *Type A* in area 94034, and one case of influenza *Type B* in 94032. Therefore, original data does not indicate epidemic outbreak but anonymized dataset might indicate outbreak of influenza (which type is unspecified). Although not shown explicitly in this example, *EI* is especially important in time series data in which same patient has to be identified over several records. This would be completely destroyed or falsified by anonymization techniques [6].

Table 2– De-identified data set with applied 3-diversity.

Name (*EI*)	Age (*QI*)	Zip Code (*QI*)	Virus (*SD*)
*	20 - 29	9403*	Flu
*	20 - 29	9403*	Flu
*	20 - 29	9403*	Flu
*	20 - 29	9403*	Flu
*	20 - 29	9403*	Flu

This leads to the conclusion that anonymization techniques solely are not sufficient. Pseudonymization techniques, on which we give a broad overview in the next section, are required and have to be explored for LPL.

Pseudonymization

In general, original dataset *D* is pseudonymized to *D'*. It consists of previously mentioned categories of data.

$$D = (EI, QI, SD, NSD)$$

In *D'* attributes are replaced with uniquely identifiable pseudonyms. Depending on use case, only *EI* have to be pseudonymized. Application of the different categories of data depends on use case. Typically, *EI* and *QI* are pseudonymized. However, *SD* and *NSD* may be pseudonymized.

$$D' = (EI', QI' SD', NSD')$$

In the following, overview of pseudonymization methods is presented.

Tokenization

Tokenization swaps distinct values with a token. Generation of the token can vary and may include a pseudo-random seed [7] or facilitate keys [8]. A distinction is made between dependent and independent tokenization. Dependent tokenization retains a relationship with original data in contrast to independent tokenization. As a result, pseudonyms based on independent tokens are more secure, because re-identification with given pseudonyms only is not possible (e.g. injection attacks) [9]. In general, generation of token is based on random seeds, cryptographic methods, or hashing [6]. An overview is given in the following.

$$Independent: TokenGenerator \rightarrow Token$$

$$Dependent: TokenGenerator(Value) \rightarrow Token$$

Random Seeds

A pseudonym can be generated based on (pseudo-) random seeds. Here, it is essential to prevent token collisions. To gain more privacy, the random seed can be combined with secret keys. An example is the patient ID generator by K. Pommerening [7].

Cryptographic Methods

Encryption approaches are either symmetric or asymmetric. Symetric approaches utilize same key for encryption and decryption. For example, Heurix-Neubauer et al. [8] and Noumeir et al. [10] utilize DES and AES. Asymmetric approaches encrypt with public key and decrypt with private key. Rottondi et al. [11] presented pseudonymization based on the RSA. The keys must be stored securely to sustain feasibility of this approach [8]. To definitively prevent de-pseudonymization the private key can be dropped. In general, the disadvantage of crypthographic methods is increased computational cost. Additionally, generated token may not be of fixed length and might get quite excessive.

Hashing

A hash function usually maps an input set to a smaller target set. Thus, most hash functions are not injective. Contrary to cryptographic methods, generated tokens have same size regardless of entrys' length. However, due to the hash function's nature, different inputs may become the same pseudonym. Therefore, collision-resistant hash algorithms are preferred for pseudonymization. Furthermore, hashing methods can be classified into keyed and non-keyed hashing. Noumeir et al. [10] and Brekne et al. [9] utilize non-keyed MDX and SHA-X algorithms. The disadvantage of non-keyed hashing is that generated tokens can be linked between different datasets if same hashing function is used. Keyed methods extend cryptographic based hashes by a key, e.g. hash message authentication code (HMAC) [12]. Assuming the key remains secret, dictionary attacks or similar attacks are prevented.

Value-Preserving Techniques

There are some further methods for tokenization, which preserve particular operations on de-identified datasets increasing utility. These algorithms are developed for specific use cases. Examples are prefix-preserving IP address anonymization [13] and distance-preserving pseudonymization [14].

Implementation Patterns

To gain additional features, like increased degree of privacy or possibility for de-pseudonymization, tokenization can be combined with additional methods. Note that the following methods comparatively perform very weak if they are used solely in a security perspective, especially for small datasets. Thus, they should always be combined with other approaches [9,15].

A *bijective mapping* stores generated token and the original value.

$$Value \leftrightarrow Token$$

Thus, it allows authorized de-pseudonymization of the token. Therefore, storage has to be encrypted to avoid unauthorized de-pseudonymization [7]. When new data entries are added to

D, which have to be pseudonymized, then each value will be looked up in the encrypted mapping store and replaced by corresponding token. If no corresponding token is found new mapping is created and appended. Therefore, *D* with less distinct values is processed faster than *D* with various distinct value. Processing performance during de-pseudonymization depends on used data structure for storage.

Limited token generation mimics original values by exclusively calculating pseudonyms with same character set and same structure. As a result, privacy increases. For instance, an input date will also result structured as date after pseudonymization [6]. *Data permutation* swaps entries based on permutation function. For each new entry, changes have to be tracked by simple mapping or by updating the function. The pseudonymization can be reverted if permutation is known [9]. *Noise addition* is used to gain more privacy by adding a pseudo-random noise to input entries [15]. However, anomaly detection or similar techniques may counter noise. *Salting* adds entropy (or salt) to token generation process to increase computing costs of insider dictionary attacks and therefore reduces risk of de-pseudonymization [12]. The salt value can either be generated deterministically from a given entry as in PBKDF2 (hashing) [12], or as a random value as in RSA-OAEP modification (encryption) [11]. For equal values salt is reused.

Reasoning for Various Pseudonymization Methods

Pseudonymization method has to be selected based on intended usage. Here, several aspects have to be considered. Beginning with properties of the data value, required utility of the token, and usage of hashing and encryption methods. The requirement for later de-pseudonymization using bijective mappings or usage one-way pseudonymization for increased security should also be taken into consideration. Moreover, multiple techniques can be added to reach the desired level of privacy/security. Due to various possibilities for pseudonymization we concluded the requirement for a general specification of pseudonymization methods within LPL and its overarching privacy framework, which we discuss in the following.

Integration in LPL

Layered Privacy Language (LPL) is intended to model privacy policies combining both legal and computer science views on privacy. The original LPL [2] has been further extended by a *UI Extension* to support privacy icons for personal privacy policy user interface [16] and *Art. 12-14 GDPR Extension* [17]. Furthermore, LPL is accommodated by an overarching privacy framework that enables *Policy-based De-identification* utilitzing anonymization methods and privacy models. To facilitate various pseudoynmization methods within LPL, both policy and *Policy-based De-identification* process have to be extended. This is shown in the following.

Pseudonymization Extension for LPL

In the following, we describe *Pseudonymization Extension* for LPL based on original formalization by Gerl et al. [2]. Note that *UI Extension* [16] and *Art. 12-14 GDPR Extension* [17] are not considered as they do not interfere with integration of pseudonymization in LPL. The root element of LPL is the *LayeredPrivacyPolicy*-element *lpp*, which remains unchanged.

$$lpp = (version, name, lang, ppURI, upp, ds, P)$$

It consists of LPL version number, privacy policies name, language defined to display descriptions, link to legal privacy policy, *UnderlyingPrivacyPolicy*-element, *DataSource*-element and a set of *Purpose*-elements.

The *Purpose*-element *p* is extended allowing definition of pseudonymization methods on specific sets of data.

$$p = (name, optOut, required, descr, DR, r, pm, D, PSM)$$

Therefore, the purpose consists of its name, status flag specifying if user can opt-out or opt-in on the purpose, a flag whether the purpose is required to be accepted, description in defined language *lang*, set of *DataRecipient*-elements *dr*, *Retention*-element *r*, *PrivacyModel*-element *pm*, and a set of *Data*-elements *d*. The set of *PseudonymizationMethod psm* is appended, which will be further detailed. Note that it is also possible to assign no *psm*, if no pseudonymization is required for the purpose.

A *PseudonymizationMethod psm* represents one pseudonymization configuration, which will be applied on the dataset.

$$psm = (name, attrName, NOD, descr, header, PSMA)$$

It is a tuple with the following attributes:

- *name*: Defines pseudonymization approach. Consists of pre-defined set of available methods.
- *attrName*: Textual representation of name for newly created attribute, which holds pseudonyms.
- *NOD*: Set of *Name*Of*Data*-elements *nod*, which represent attributes to be pseudonymized. It refers the *name* attribute of the *Data*-element. It must at least consist of one valid *name*.
- *descr*: Human-readable description of pseudonymization in defined language *lang*.
- *header*: Human-readable header of pseudonymization in defined language *lang*.
- *PSMA*: Set of *PseudonymizationMethodAttribute*-elements *psma* describing further configurations of pseudonymization approach.

Each *PseudonymizationMethodAttribute*-element *psma* configures the defined approach:

$$psma = (key; value)$$

This key-value tuple defines attributes necessary for all possible methods.

LPL has been extended by adding *PseudonymizationMethod*-element *psm*, and *PseudonymizationMethodAttribute*-element *psma*, which allows definition of various pseudonymization methods (see Figure 1). In the current state, the extension supports various hashing approaches (SHA-X, MDX), keyed hashing (HMAC SHA-X), keyed hashing with entropies and mapping (PBKDF2 HMAC SHA-X), symmetric cryptography (AES, DES, 3DES, RC4, Blowfish), and random seeds with mapping. Further approaches can easily be added by defining them as *psm* and *psma* elements.

Secrets, which can lead to re-identification, are not stored in the LPL model. According to GDPR, these information must be stored separately [Recital 29, 1]. The concrete key and secret management belongs to the controller, entity which administrates stored data.

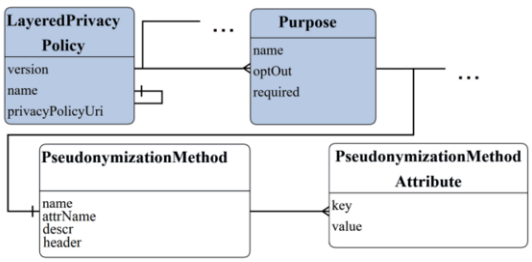

Figure 1– LPL structure extended by the pseudonymization elements. Further elements and attributes are omitted for the scope of this paper.

Policy-based De-identification

Policy-based De-identification based on LPL requires requesting entity to provide the purpose, a set of data , and set of data sources for the request. The requesting entity will then be authenticated. The purpose, entity and set of data will be authorized against properties of respective privacy policies *lpp* of the data sources. During *Minimum Anonymization* personalized privacy settings of each data source will be applied. Lastly, a common privacy model is derived from all related *lpp* and applied on the dataset.

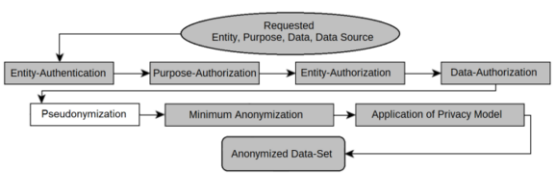

Figure 2– Modified Policy-based De-identification processes of the LPL framework when data is requested [2].

This process chain will be extended by adding the *Pseudonymization* process before *Minimum Anonymization* (see Figure 2). An inverted order, anonymization before pseudonymization, would result in less utility preservation, due to possible anonymization of values that should be preserved by tokenization with bijective mapping. Each pseudonymized *Data*-element will be altered after pseudonymization. For example, classification will be changed to *NDS* to avoid any further anonymization by later processes.

Health Care Scenario Results

With pseudonymization extension we can now de-identify the influenza dataset (see Table 1) and simultaneously add possibility of re-identification using bijective mappings. Patient names and virus type in the influenza example will both be stored as pseudonyms or IDs. For instance, if new insights regarding patients' health arise, IDs can be de-pseudonymized to deliver these new information to affected patients by gaining their identity and virus information.

The following configuration, using HMAC-SHA-1 algorithm and patient ID generator [7] combined with mappings, would achive such pseudonymization:

psm_0 = ("HMAC-SHA-1", "Name -ID", {"Name"}, "Description_0", "HMAC-SHA-1 pseudonymization", {})

psm_1 = ("PID", "Virus-ID", {"Virus"}, "Description_1", "PID pseudonymization", {})

It is noteworthy that the key, necessary for HMAC, and mapping are not defined within the LPL privacy policy to comply with GDPR [Recital 29, 1].

The results of the pseudonymization, applied on the raw data (see Table 1), defined by *psm_0* and *psm_1* are given in Table 3, which is accessible by the data analyist. Note that in Table 3 anonymization or privacy models are not applied. The mappings, necessary for re-identification, can be examined in Table 4, which should be secured and only be accessible by authorized personal. The extension also supports an AES encryption for mappings to increase security properties.

Table 3– De-identified dataset with pseudonymization. Pseudonyms shortened for better readability.

Name-ID (*NSD*)	Age (*QI*)	Zip Code (*QI*)	Virus-ID (*NSD*)
EA4B255	28	94032	KM93N2O
ADE0D85	25	94032	KM93N2O
2412F8F	22	94034	93I8M72
FF85768	20	94032	0O9INMW
71624DB	24	94034	KM93N2O

Table 4– Mappings which can restore content replaced by pseudonyms.

Name-ID	Name
EA4B255	John
ADE0D85	Max
2412F8F	Mary
FF85768	Harry
71624DB	Theresa

Virus-ID	Virus
93I8M72	Flu Type A
0O9INMW	Flu Type B
KM93N2O	Flu Type C

If we now compare anonymized data (see Table 2) and pseudonymized data (see Table 3), faulty conclusion cannot be made anymore. It can be observed that different types of viruses are present, which can uniquely be assigned to individuals. Still it can be seen that there is clustering. But with bijective mappings it can be shown that often appearing virus *KM93N2O* is the relatively harmless influenza *Type C*. Therefore, conclusion that there is no influenza epidemic, can be drawn. Furthermore, pseudonymization of SD can mitigate *attribute linkage* [4], identification by sensitive attribute. Due to pseudonymization, the attacker cannot infer privacy by identifying individuals by unique sensitive attributes or utilize external knowledge because provided information is not readable by the attacker.

Discussion

Based on verification of pseudonymization extension of LPL against introduced health care scenario, we conclude that anonymization methods and privacy models are insufficient for the given health care scenario. This is also valid for other anonymization techniques and privacy models. The application of pseudonymization methods tackles presented flaws by preserving essential data utility which would be lost otherwise. This has been exemplarily validated on the given scenario. To the extent of our knowledge, LPL is the first privacy language that models and enforces pseudonymization methods. Implementation is currently limited and in the state of work-in-progress. However, potentially various pseudonymization methods (which follow described configuration structure) may be added. For instance, asymmetric cryptographic functions and non-keyed hashing approaches with use of mappings for re-identification are appropriate additions. LPL certainly enables arbitrary combinations of pseudonymization and

anonymization. It is however important to choose suitable methods (pseudonymization, anonymization, privacy models) based on intended purpose, otherwise privacy may be broken. The selection of appropriate de-identification should be executed by an expert, e.g. trained data protection officer.

Conclusions

Privacy is an emerging topic which has to be considered within various fields. In health care it is especially necessary that data privacy is considered because processed data falls under the special categories of data. LPL intends to model privacy policies to inform *Data Subjects* and enforce privacy by design. In LPL, solely anonymization and privacy models were defined and used for de-identification. However, it was shown that pseudonymization is necessary for health care scenarios. *Pseudonymization Extension* of LPL offers possibility to define pseudonymization methods that have to be applied on the dataset. For example bijective mappings would allow an authorized de-identification, which is not possible for anonymization only techniques. Additionally, tokenization allows clear distinction of original values. *Policy-based De-identification* process joins pseudonymization, anonymization, and privacy models as a holistic approach. LPL and its overarching framework are continiously developed and researched. Evaluation of *Policy-based De-identification* process based on real health care data is sort after. Also, evaluation of LPL in other domains like *IoT*, *Cloud,* and *Mobility* are aimed for. The framework itself should be extended for fullfilment of *Data Subject Rights* to support *Data Protection Officers*. Suitable user interfaces for different domains and user groups (e.g. elderly or children) are developed to allow GDPR-compliant consent.

References

[1] Regulation (EU) 2016/679 of the European parliament and of the council (General Data Protection Regulation), European parliament and council, April 27, 2016.

[2] A. Gerl, N. Bennani, H. Kosch, and L. Brunie, LPL, towards a GDPR-compliant privacy language: formal definition and usage, In: *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVII*, Springer, 2018, 41-80.

[3] P. Samarati and L. Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, In: *Technical Report*, SRI International, 1998, 101-132.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, *ACM Trans Knowl Discov Data* **1** (2007), 3.

[5] C. Dwork. Differential privacy: a survey of results, In: M. Agrawal, D. Du, Z. Duan, and A. Li (Eds), *Theory and Applications of Models of Computation*, Springer-Verlag, Berlin, Heidelberg, 2008, 1-19.

[6] N. Venkataramanan and A. Shriram, *Data Privacy: Principles and Practice*, Chapman and Hall/CRC, 2016.

[7] M. Lablans, A. Borg, and F. Ückert, A RESTful interface to pseudonymization services in modern web applications, *BMC Med Inform Decis Mak* **15** (2015), 2.

[8] J. Heurix and T. Neubauer, Privacy-preserving storage and access of medical data through pseudonymization and encryption, In: *International Conference on Trust, Privacy and Security in Digital Business*, Springer, Berlin, Heidelberg, 2011, 186-197.

[9] T. Brekne, A. Årnes, and A. Øslebø, Anonymization of ip traffic monitoring data: attacks on two prefix-preserving anonymization schemes and some proposed remedies, In: *International Workshop on Privacy Enhancing Technologies*, Springer, Berlin, Heidelberg, 2005, 179-196.

[10] R. Noumeir, A. Lemay, and J.M. Lina, Pseudonymization of radiology data for research purposes, *J Digit Imaging* **20** (2007), 284-295.

[11] C. Rottondi, G. Mauri, and G. Verticale, A data pseudonymization protocol for smart grids, In: *Online Conference on Green Communications (GreenCom), 2012*, IEEE, 2012, 68-73.

[12] H. Aamot, C.D. Kohl, D. Richter, and P. Knaup-Gregori, Pseudonymization of patient identifiers for translational research, *BMC Med Inform Decis Mak* **13** (2013), 75.

[13] J. Fan, J. Xu, M.H. Ammar, and S.B. Moon, Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme, *Computer Networks* **46** (2004), 253-272.

[14] F. Kerschbaum, Distance-preserving pseudonymization for timestamps and spatial data, In: *Proceedings of the 2007 ACM workshop on Privacy in electronic society*, ACM, 2007, 68-71.

[15] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Technique (WP216), Data Protection Working Party, April 10, 2014.

[16] A. Gerl, Extending layered privacy language to support privacy icons for a personal privacy policy user interface, In: *Proceedings of the 32nd International BCS Human Computer Interaction Conference*., Belfast, Ireland, 2018, 1-5.

[17] A. Gerl and D. Pohl, Critical analysis of LPL according to articles 12-14 of the GDPR, In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*, ACM, Hamburg, Germany, 2018, 1-9.