INTED2020

14th annual International Technology, Education and Development Conference
Valencia (Spain), 2 – 4 March, 2020

THREE-LAYER MODEL FOR LEARNER DATA ANONYMIZATION

A. Aleksieva-Petrova[1], I. Chenchev[2] and M. Petrov[3]

[1]Technical University of Sofia, Faculty of Computer System and Technology (BULGARIA)
[2]Technical University of Sofia, Faculty of Computer System and Technology (BULGARIA)
[3]Sofia University "St. Kliment Ohridski", Faculty of Mathematics and Informatics (BULGARIA)

INTRODUCTION

Nowadays the increased use of learning management systems (LMS) leads to process large amounts of data for educational or administrative purposes. The data can be stored into database(s) and associated services could retrieve it and analyse it. Most of this data is sensitive in terms of privacy. The EU General Data Protection Regulation (GDPR) requires companies, organizations and/or any individuals to comply with it for better protection of an individual personal data while ensuring that the individual data is trusted, secured and governed..

To address such problems, we propose a three-layer method for anonymization, which consists of three layers: Privacy Compliance Layer, Data Anonymization Layer and Analysis & Reporting Layer. The proposed method used different sources of data in field of LMS as experimental data in order to achieve learning analytics.

LEARNER RELATED INFORMATION TOPOLOGY

Personal data includes a learner's name, an address, a date and/or place of birth, social security numbers, student-identification numbers, race, gender, economic status, some digital files such as photographs and other forms of information that may reveal a specific learner's identity.

The second data category is related with academic information (such as the educational organizations which a student attends, courses, growth, enrollment, grades, completion, etc.) and various other forms of data collected for learning experience including evidence of learning outcomes (formal and informal) and learning activities (attendance, behavior, extracurricular activities, programs participation, etc.).

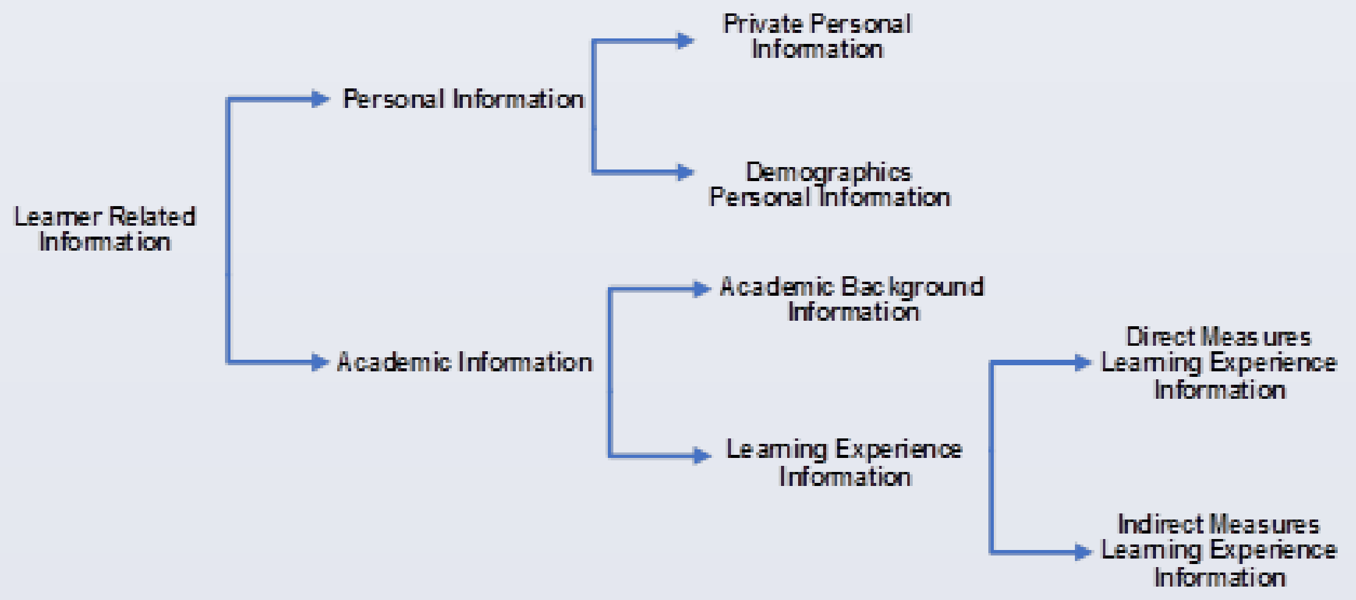
Learner Related Information
Personal Information
Private Personal Information
Demographics Personal Information
Academic Information
Academic Background Information
Learning Experience Information
Direct Measures Learning Experience Information
Indirect Measures Learning Experience Information

Figure 1. Learner Related Information Topology.

ANONYMIZATION

Few common types of data anonymization are the following:

1) Removal
2) Modification
3) Encryption
4) Hashing
5) Data Masking
6) Initial Systems' Design for Data Privacy and Security

We propose the modification of hashing called Squeezed Hashing. To make it theoretically and practically impossible to recover the original data, some of the bits from the generated message digest to be omitted – for example, first 4 even bits and last 4 even bits from the digest can be deleted (skipped). With that way of generation, the message digest will be the same for the same input data. And also – for the analysis and reporting purposes the collision resistance of the hash algorithm is not important (if any at all).

THE PROPOSED 3-LAYER METHOD FOR ANONYMIZATION

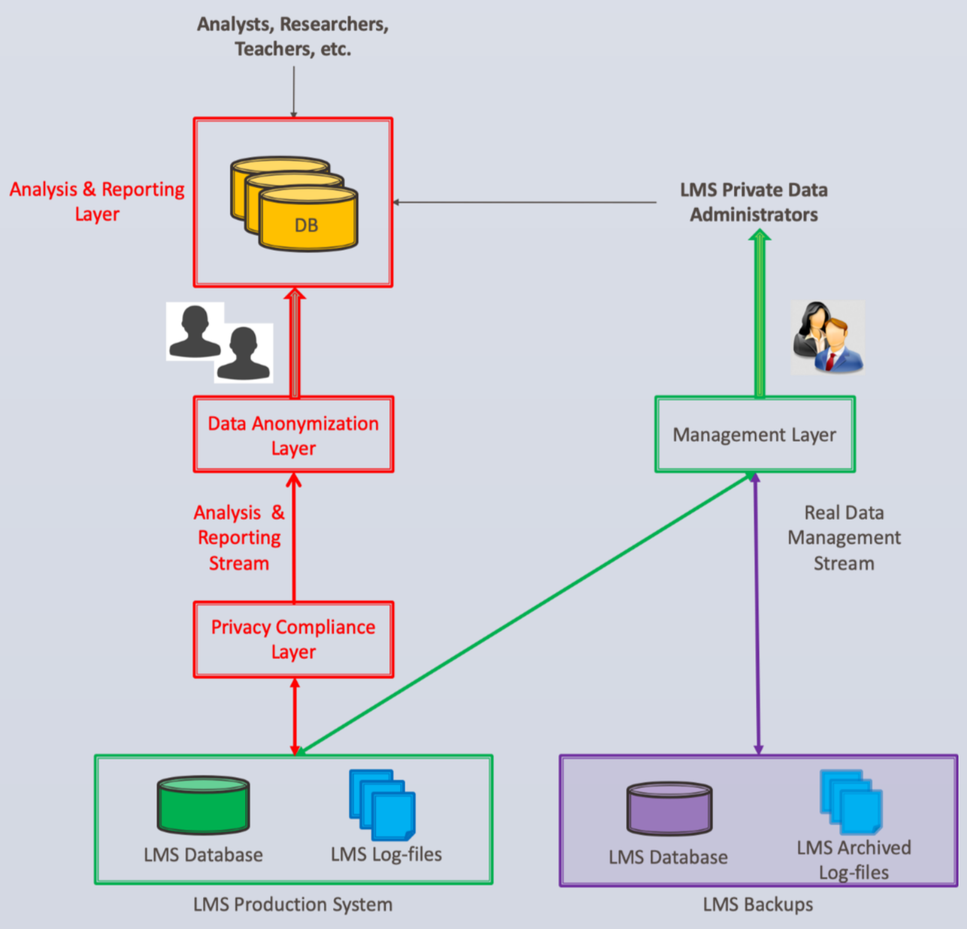

Figure 2. The 3-Layer Model for Data Anonymization.

The structure of the data model is based in a three-layer model for anonymization:

Layer 1 – Privacy Compliance Layer

Layer 2 – Data Anonymization Layer

Layer 3 – Analysis and Reporting Layer.

Layer 1 is the "Privacy Compliance Layer"

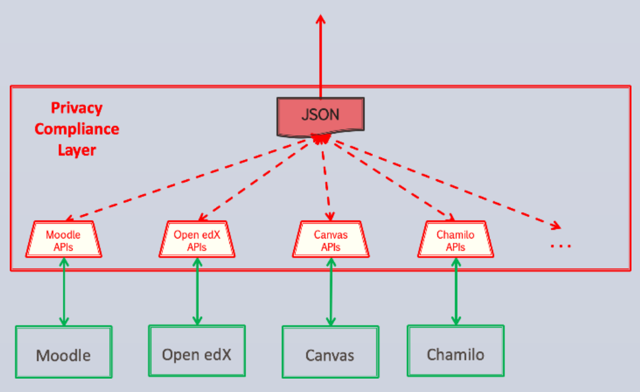

Figure 3. The "Privacy Compliance" Layer.

Once, the information is retrieved from the LMS system, it will be transformed into a file in JSON format. The JSON file has four attributes: Data Field (contains the value of the original attribute), Name Field (the name of the original attribute), Field Type (integer, string, blob, etc.), Secret Field (information is secret or not secret) – in case of a "secret" field(s) -> the data section has to be anonymized.

[{	
"Name":	"First.Name.Of.Student",
"Secret":	"hash",
"Type":	"string",
"Values": }, {	"Ivaylo"
"Name":	"Last.Name.Of.Student",
"Secret":	"sqhash",
"Type":	"string",
"Values": }, {	"Chenchev"
"Name":	"Student.Age",
"Secret":	"mask",
"Type":	"integer",
"Values": },	24
]	

Figure 4. Example of structure of JSON file format for "Data Anonymization" Layer.

The field "Secret" contains the following values: ["hash" | "sqhash" | "masking" | "false"].

If it is set to "hash" – then the contents of the "Value" attribute will be anonymized with one-way hash function.

If it is set to "sqhash" – then the contents of the "Value" attribute will be anonymized with one-way hash function and 8 bits will be removed (as proposed in the previous section with the "squeezed hash" anonymization). In the above example the student's Family name will be anonymized with "squeezed hash" method.

If it is set to "mask" – then the contents of the "Value" attribute will be anonymized with masking (the number will remain as number).

If the "Secret" field is set to "false" – then the "Value" attribute will not be changed.

Layer 2 is the "Data Anonymization" Layer

This is the core layer from our three-layer model. Here, the collected information from previous layer will be anonymized. This process is based on the retrieved data and its type (received from previous layer). It is important to guarantee the mapping of input data and the generated "anonymized" data.

[{	
"Name":	"First.Name.Of.Student",
"Secret":	"hash",
"Type":	"string",
"Values":	"2261ace9b7cdaa96d4980f9b08290f70de96ad769cc0677d8762208eaae469e8"
	. . .
{	
"Name":	"Student.Age",
"Type":	"integer",
"Values": },	25
]	

Figure. 6 Example of applied method.

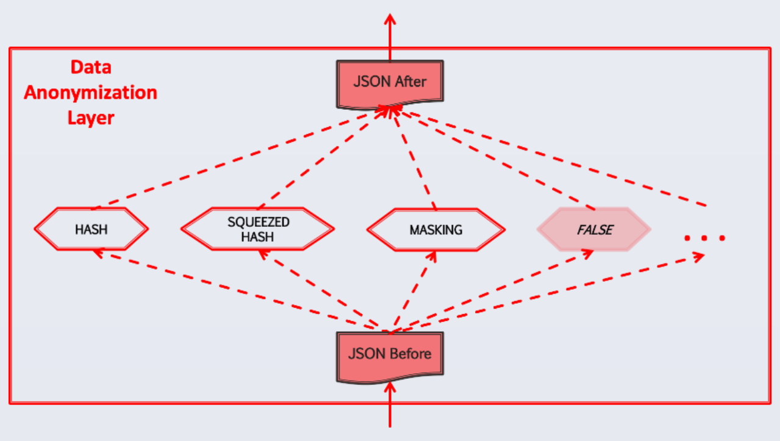

Figure 7. The "Data Anonymization" Layer.

Layer 3 is the "Analysis & Reporting Layer"

This layer will keep and store the previously collected and already anonymized data. It provides the data to any teachers, content writers, analysts, researches, etc., interested in analyzing it.

It can be kept in the JSON files (after the data anonymization) and/or it can be stored in databases. Each data under GDPR regulation and has value "True" in field "Secret" could be store separated from those how has value "False".

In order to be able to uniquely identify the data to which the individual refers every moment, it is advisable to generate a learner identifier. It will be used as pointer to Personal Information Identifier and Academic Information Identifier.

CONCLUSION AND FUTURE WORK

All LMS and learning services used in education collect different types of learner data. One type of data is related to the personal data, and another to the learning achievements in different fields. Over time, this data increases and its sharing and analysis raises a major problem with the protection of personal data. We present our approach, called three-layer model, to try to solve the problem with learner data anonymization. With the proposed model and methodology for anonymization is kept the uniqueness of the data entities with losing the personally identifiable information.

To describe further and illustrate our proposed approach, a system architecture and a prototype will be designed and implemented.

ACKNOWLEDGEMENTS

The research in this paper was supported by Contract KP-06-OPR/1 from 13.12.2018 project "An innovative software platform for big data learning and gaming analytics for an user-centric adaptation of technology enhanced learning (APTITUDE)", Competition for financial support of basic research projects on societal challenges – 2018, Bulgarian National Science Fund, Ministry of Education and Science, Bulgaria.

CONTACTS

Assoc. Prof. Adelina Aleksieva-Petrova, PhD
Technical University of Sofia, Bulgaria,
Faculty of Computer Systems and Technologies

Telephone: (+ 359) 877 706 778
E-mail: aaleksieva@tu-sofia.bg
adelina.aleksieva@gmail.com

Ivaylo Chenchev, MSc
Technical University of Sofia, Bulgaria,
Faculty of Computer Systems and Technologies

Telephone: (+ 359) 888 500 509
E-mail: ivaylo.chenchev@gmail.com

Assoc. Prof. Milen Petrov, PhD
Sofia University "St. Kliment Ohridski",
Faculty of Mathematics and Informatics,
Department of Software Engineering, Bulgaria

Telephone: (+359) 878 594 220
E-mail: milenp@fmi.uni-sofia.bg
milen.petrov@gmail.com