LMS DATA COLLECTION, PROCESSING AND COMPLIANCE WITH EU GDPR

A. Aleksieva-Petrova¹, I. Chenchev¹, M. Petrov²

¹Technical University of Sofia (BULGARIA)
²Sofia University (BULGARIA)

Abstract

Nowadays, the learning process is supported by different Information and Communication Technology (ICT) tools which provide more flexible methods of delivery of learning content and students’ evaluation. The Learning Management Systems (LMS) are used to integrate a wide range of pedagogical and course administration tools.

The most widely used LMS system is Moodle. Its integration with different tools improves and supports the learning process. Key component of the system is the course, which can be organized in different formats and can include different activities and resources. Students are being enrolled to a specific course. After a successful course delivery, there exist well-defined assessment methods, which give a lot of opportunities. One of the faced challenges is to apply different learning analytics methods in order to take decisions and conclusions for course content from one side and students results, and motivation from another. For that purpose, personal information is stored and processed.

The main goal of the paper is to review the GDPR regulation and to apply it to a life LMS system (Moodle). There exist some plug-ins for managing and processing personal data within Moodle system. The research here focuses on life Moodle system and possible places (outside of the system itself), where personal information can be stored. A chapter five describes the authors approach for data processing with two processes - for data anonymization and for data sharing with external systems.

Keywords: LMS, GDPR, Moodle, private information, compliance, database, log-files, analyse, protect.

1 INTRODUCTION

Learning Management Systems (LMS) are enterprise-wide and internet-based systems, which integrate a wide range of pedagogical and course administration tools. Authors in [2] are talking about a significant change that takes place in higher education. It is related to the integrated computer systems known as LMS. They allow instructors and students to share course materials, make class announcements, submit and return course assignments, and communicate with each other online [6]. Some researches investigates software architecture for production and delivery of learning resources with audio elements in university programming courses [12] or design and implementation of learning platform ARCADE [13]. Other study create Moodle’s Web Service API, which wraps up the retrieving of course details and resources, in order to deliver all available resources of a given course keeping the internal structure of course organization [14].

A detailed overview of the rapid evolution of online LMS and a comprehensive review of their influence on the teaching and learning is presented in this paper. We are sharing their concerns about possible corporatization of academic knowledge and the importance of steps that have to be taken to identify how LMS systems can be used to augment and complement an institution’s core teaching objectives. Therefore, the collecting of the statistical and analytical data becomes very important.

One of the most widely used LMS is the Moodle system [5]. The results of the survey in Angola High polytechnic school show that at the third place teachers and students recognize LMS as Information and Communication Technologies (ICT) tool [11]. It processes few types of data – the online data like data for users, courses, assignments, results and real-time activities. The both of those types are stored in a database. Together with the general information – like course materials, a personal information is also collected and stored in the Moodle’s database. It is also stored in the database backups, in the snapshots of the system and in the different archives.

Moodle processes different types of data stored in tables, logically grouped in two major types: information for users and courses, and activities. The latest version (at the time of paper preparation is 3.6.3+, released on date 2019-05-03) of Moodle database has 446 tables (with eleven plug-ins like VPL,
Questionnaire etc.). This is an upgraded version (not new installation), because the research was performed in a life Moodle system. Only thirty-one of them contain the word ‘user’ in theirs name and another fifteen tables have the word ‘log’ in theirs names. There already exist some plug-ins for Moodle’s compliance with the GDPR regulation.

The paper is organized as follows: in next section – the problem of GDPR compliance is introduced and analyses the possible targets, where personal information can be stored and extends the understanding of being compliant. Section three gives some highlights of the GDPR and specifics about the personal information internals. Section four represents some high-level details of the LMS database structure (in regard to Moodle), related to user’s information and real-time activities logging. It also summarizes the different types of data, stored in Moodle. In section five data processing is summarized as regulated in GDPR. Section six gives an extended overview of what should be taken into consideration, when discussing whole software system compliance. The last section gives ideas for further research and specific conclusions.

2 GENERAL DATA PROTECTION REGULATION

In [1] is provided a framework to help understand the basic principles of the GDPR, focusing on the data collection and processing. Collection and processing of personal information is fundamental for the work of researchers. So, data protection of collected and processed data is of the highest importance.

This Regulation was published in the Official Journal on 4 May 2016 and became active on 25 May 2016, but the most essential provisions became applicable in all Member States from 25 May 2018 [8].

The GDPR introduces two obligations as two roles - from one hand are the "controllers" and from another – "processors" [3]. A controller role regulates all of the purposes and means of processing personal data while processor is responsible for processing data on behalf of the controller. A processor is a person who acts on the controller’s instructions.

The GDPR considerably increases the range of the regulatory compliance for organizations which process data on behalf of the data controllers – so-called "data processors" [7]. Data processors are required to implement any appropriate security measure, report data breaches to the controller, keep a register of the data processing activities, and apply controller’s authorization before delegates or redirects to the third-party software or services to further process personal data. As stated in [4], the processors are also directly responsible to implement mechanisms for failure detection, recovery and protection to comply with the GDPR. Further in [4] can be found the systematic representation of the general processing principles. The key responsibility under the GDPR is to comply with all six general processing principles when processing personal data:

1. Lawfulness, Fairness and Transparency - personal data must be processed lawfully, fairly and in a transparent manner;
2. Purpose Limitation – personal data must be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes;
3. Data Minimization – personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed;
4. Accuracy – personal data must be accurate and, where necessary, kept up to date. Inaccurate personal data should be corrected or deleted;
5. Retention – personal data should be kept in an identifiable format for no longer than is necessary;
6. Integrity and Confidentiality – personal data should be kept secure.

Personal data can be defined as information, which is related to identify or identifiable live person (human being). It is a broad term and includes a wide range of information. The GDPR expressly states personal data includes online identifiers (such as an IP addresses and cookie identifiers). However, this is already likely to have been the case under the Data Protection Directive - Breyer C-582/14.

The data subject is the live person to whom the personal data relates. Processing is a very broad concept and includes almost anything you can do with personal data, including collection, use, storage and releasing/destroying.

The GDPR applies the same broad security obligation as the old Data Protection Directive, requiring controllers and processors to take appropriate technical and organizational measures to protect their
systems. This broad obligation is supplemented by additional obligations to take the following steps, where appropriate:

- the pseudonymisation (pseudo-anonymization) and encryption of personal data;
- the ability to ensure the ongoing confidentiality, integrity, availability and resilience of its information technology systems;
- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident; and
- Process for regularly testing, assessing and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing.

3 LMS DATA STRUCTURE ANALYSIS

First for the data structure analysis we use following system settings: Moodle LMS version is 3.6.3+ (Build: 20190503).

3.1 Structure of Moodle System - Database

We use database server version 10.1.26-MariaDB-0+deb9u1 Debian 9.1. The number of tables is 446, where there are 31 tables with string “user” in theirs name and no tables with string “student” and “teacher” in theirs name. There are 15 tables with string “log” in theirs name.

The table user has 53 columns. It contains one record for each person register in the system. It plays main role in database and has relations to most of the tables in Moodle database (more than half of all tables in the database). The table user_devices stores user’s mobile devices information in order to send PUSH notifications. If you want to have information for who is participating/visible in the course, you can see this information from user_enrolments table. The information for customizable fields categories, user fields and user profile fields are received by follow tables: user_info_category, user_info_data, user_info_field. In order to keep track of course page access times, used in online participants block, and participants list the user_lastaccess table is used. The other tables are: user_password_history for a rotating log of hashes of previously used passwords for each user; user_password_resets table for tracking password reset confirmation tokens; user_preferences to store arbitrary user preferences and user_private_key table to access keys used in cookieless scripts - rss, etc.

There are many other tables that play a key role in the system and contain information that is needed for courses and different types of activities such as forum, chat, assignments, glossary, book, wiki, etc. Moodle DB design is good example for implementation and relationships between users and other tables - established with foreign keys.

According to the process of system backups there are three main DB tables related to it: backup_controllers, backup_logs, and backup_courses.

One of the main resources for personal data storage are the logs in Moodle, which are nothing more but tables, filled with actual student’s activities. Logs are available at site level and at course level. A log of activity in the course may have any combination of group, student, date, activity, actions and level. The filter can be applied at teaching level r at participating level. The first one is an event or action performed by a teacher which affects the students’ learning experience. The participating level is an event or action which could be related to a user’s learning experience. For example, log of activity in the course has direct identification as user full name and indirect identification as user id and IP address. A log of site activity can display all activities, site news or site errors.

The Moodle system has implementation of GDPR principles, included in the standard distribution of Moodle 3.5 as two plugins: policies plugin and data privacy plugin. The policies plugin allows definition of various policy documents (site policy, privacy policy, intellectual property policy, late assignments policy and others) and provide a new user sign-on process, with ability to define multiple policies (site, privacy, third party), track user consents, and manage updates and versioning of the policies [9].

The Data privacy functionality provides the workflow for users to submit subject access and erasure requests and for site administrators and privacy officers to process these requests. It also includes the data registry to define a purpose and retention period for data stored in a Moodle site [10].
3.2 Structure of Moodle System – Log levels

There are six levels of logs:
- Logs of course activity
- Levels
- Logs of site activity
- Live logs
- Site administration settings
- View logs capabilities

In the Moodle version, which was used for the tests (Moodle 3.6.3+ (Build: 20190503)), there are three possibilities to log data (called 'log stores'):

1. Standard log: This is the default possibility. All the log data will be saved to the database table 'logstore_standard_log'. Here for example, can be seen information when a user has viewed a particular course, when a user has logged into the system, or logged out, etc.

2. Legacy log: This is a log structure used in earlier versions of Moodle. All log data have been saved to the database table called 'log'.

3. External database: there is the possibility to select and configure an external database in order to store log data. The database connection, name, and other settings can be configured under [Site administration] -> [Plugins] -> [Logging] -> [Manage log stores] -> [External database log settings].

3.3 Data types categorization analysis, based on Moodle LMS

One of the basic principles underlying the GDPR is to collect the requested information from individuals and use it only for limited purposes. The LMS data can be divided into two types - structured and non-structured data. Table 1 summarizes what type of personal information is discovered in the different types of data sources. Structured data is divided into user data and activities data. Non-structured data is defined as data from uploaded external resources in different formats - such as text files, images, video, audio and others. Both structured and unstructured data is categorized as data that can identify user directly (or automatic – with the available data) or indirectly (that requires extra data manipulation).

According to [5], personal information is the one that is used to identify someone. The direct identification, as defined in previous section, can be used to view the information that is stored for each user, such as: firstname, lastname, email, picture, address, e-mail address, social network profiles (icq, skype, yahoo, etc.).

For indirect identification can be considered all records that are related to stored information about individual users' activities, such as IP address, user names, interest areas, and so on.

Table 1. LMS data types categorization – direct and indirect identification.

<table>
<thead>
<tr>
<th>LMS Data</th>
<th>Structured Data</th>
<th>Non-Structured Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Users (Database)</td>
<td>Activities Logging (Database)</td>
</tr>
<tr>
<td></td>
<td>Direct identification</td>
<td>y</td>
</tr>
<tr>
<td></td>
<td>Indirect identification</td>
<td>y</td>
</tr>
</tbody>
</table>

As can be seen from the table – structured data can lead to easy identification of the user, while unstructured data needs some extra (pre)processing in order to identify user and makes identification difficult.
4 LMS SYSTEMS COMPLIANCE WITH GDPR

In order to define methodology of GDPR principles in the LMS on the fig. 1 we define a hierarchical model of the different locations, where personal data can reside for a given LMS system.

In a life LMS system (for the purpose of this research is the Moodle system), all of those places/locations can contain personal information. Obviously, it is not enough to develop a plug-in(s) to only manage the Moodle’s database. For a LMS system, to comply with the GDPR regulation is necessary to extend the scope.

In order to define the GDPR personal data in LMS, it is necessary to make review of the answers, given to number of questions.

- **The first question**: Why it is necessary to held personal information in LMS?

  In LMS, the main roles of users are teachers and students. In order to be able to implement and effectively use such systems in the learning process on both sides, it is necessary to store information that identifies the user. The teacher should be able to create and deliver learning content, as well as organize different types of activities and assessments for students. Students themselves need to have access to the resources provided, to participate in the tasks assignments and to be evaluated.

- **The second key question**: How to store the information?

  As a source of personal information, different types of structured and unstructured resources can be defined. Structured resources include table records that build the database, their relations, and log files of an LMS. The other sources of information about personal data is the resources which are in use when creating the content of the course or the students’ projects. These are typically different types of multimedia resources such as text, image, video, audio, and etc. They are usually stored as files in various extensions, such as pdf, doc, ppt, mp3, mp4, and etc. In these resources, it is possible to implicitly store a direct identification of the individual.

- **A third major question** (relates to the problem of) Who can access to those data?

  The main users in an LMS are lectors, students and administrators. Different types of lectors are defined in one of the systems, depending on the limitations that are set in terms of access to the courses and their management. For example, Moodle defines roles as manager, course creator, lecturer, non-editing teacher, and others. Administrators can also be defined at different levels of permissions depending on their system resource management rights. Typically, this kind of systems provide a flexible scheme to define new roles and set their permissions in the system. It is therefore necessary to regulate and respect this process with great care.

It is necessary to take into account when the personal data will be deleted or eliminated. This should be the right of every user to ask for his / her data to be deleted. Then two basic questions arise. The first sub-question is to delete the data from the archives and the logs. The second key sub-question is how this will affect the process of student learning and assessment.
As result in next section of paper we propose model of data anonymization, using results of analysis of answers of questions above.

5 LMS DATA PROCESSING APPROACH

In previous section we make categorization of data types, and in current section we propose how to process data and anonymize data, in order to be compliant with the user requirements and settings and how/when to share data with external processor tools.

First, the process is divided into two sub-processes – data anonymization process – with four steps (and one pre-step), and data-anonymization sharing for processing by external tools – in two steps.

Data anonymization process (as depicted on fig.2) contains following steps:

- Step 1 - initial data
- Step 2 - User data (filter by policy) – directly identifiable, User 'data' (filtered by preferences) – indirectly identifiable,
- Step 3 - Enforce anonymization usage, and
- Step 4 – Anonymous data – with their usage attributes (what can be done and how can be shared). In last step – data are already filtered in previous steps and contains only valid for data processing, and how these data are related for processing with (possible) external systems and tools.

![Diagram](image)

Figure 2. Proposed process of anonymization.

Second sub-process is called data anonymization sharing. After completing the first process of data filtering and tagging with shared attributes – the data can/cannot be shared with external processors (as depicted on fig. 3).

In this scenario from previous sub-process have four use cases:

1. The user does not agree to share anonymous data with external tool or with specific tool - in this case we cannot apply sharing anonymous data.
2. The user agrees to share anonymous data, but no back-track possible.
3. The user agrees to share data and in very strictly controlled cases agrees to be possible to de-anonymize data (deanonymization for external tool or deanonymization for internal system, without giving feedback to external tool).
4. Possible de-anonymization on request by external tool.
External processor can process anonymous data, can request de-anonymization (which can be automatically denied by internal system), and works with eventually anonymous data. We define *eventually anonymous data*, which can be easily de-anonymized, but de-anonymization process is never applied.

6 CONCLUSIONS AND FUTURE WORK

The following general conclusions can be summarized by the research done in current paper:

- High-level overview of General Data Protection Regulation;
- Classified LMS Data Models and Structures, according to the GDPR;
- Defined hierarchical structure model of LMS Systems compliance with GDPR;
- Proposed process modelling and approach for anonymization and LMS data processing, which can be generalized to any information system.

Definitely, the area for Moodle’s compliance with GDPR is still to be explored. The following points highlights three possible areas of further research:

1. Identification and preparation of proper backup/restore strategy for Moodle system in compliance with the GDPR regulation;
2. Preparation of methodology for automatic data anonymization - in order to collect statistical and analytical data, such as usage of the system, usage of the courses, behavior of the students in learning process of different areas, reports and etc. for further research.
3. Define a methodology for GDPR compliance in the LMS systems in general – from the initial systems architecture design, to the data collection, data processing, storing and releasing/deleting the personal data.
4. Validate proposed approach for Anonymization and LMS Data processing with real example.

Points for improvements and further research gives starting point for different directions of research, nevertheless we believe that proposed classification, models and approaches are significant step to creation of software platform for analyzing of learning and game data for a user-oriented adaptation of technology-assisted training by providing framework for anonymization. Anonymization is crucial point to system that analyses big data, related to any kind of users.
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.

REFERENCES


[8] https://eugdpr.org/


