The 12th International Scientific Conference
eLearning and Software for Education
Bucharest, April 21-22, 2016
10.12753/2066-026X-15-178

RELATIONS BETWEEN LEARNING ANALYTICS AND DATA PRIVACY IN MOOCs

Malinka IVANOVA
Technical University of Sofia, Bulgaria
m_ivanova@tu-sofia.bg

Carmen HOLOTESCU
„Ioan Slavici” University of Timisoara, Romania
carmenholotescu@gmail.com

Gabriela GROSSECK
West University of Timisoara, Romania
ggrosseck@e-uvt.ro

Cătălin IAPĂ
University Politehnica Timisoara, Romania
catalin.iapa@gmail.com

***Abstract**: Massive Open Online Courses (MOOCs), as a new educational trend, attract worldwide a huge number of learners giving them appropriate knowledge according to personal needs. Learners are stimulated to be active participants and facilitated to create distributed learning communities through: sharing opinions and setting preferences; analyzing, annotating and translating learning materials; participating in interactive gaming exercises; communicating with other participants, tutors and guest lectures; preparing and submitting practical assignments; group working and peer rewieving; proposing other useful (open) educational resources; publishing insights on different Social Media platforms. Therefore, during participating in Massive Open Online Courses a wide variety of students' personal and activity data are collected and used for different purposes by tutors, hosting platforms and possible third party actors, such as universities or companies. Part of this big data is processed by the Learning Analytics modules, in order to understand and optimize the learning process. The questions that arise are: - What kind of students' private data are shared and why?, - What are the privacy models provided by MOOCs?, - What data is needed for Learning Analytics modules?, - Are there any third party actors processing the students' data and why?, - Which are the relations between Learning Analytics and Data Privacy?, - Is it possible for the learners to be protected?. The paper is looking for answers of the above placed questions summarizing and analyzing the current projects and practices, best learning cases and research standpoints. The findings are used for development of a model presenting a possibility to ensure data privacy in MOOCs. The analysis, model and conclusions indicate the positive features and bandgaps in the area of data privacy in the context of MOOCs. They could be used in the form of directions for developers of MOOCs with aim to guarantee data privacy of learners.*

***Keywords**: MOOCs; Learning Analytics; Data Privacy; Big Data; Higher Education.*

I. INTRODUCTION

Learning Analytics is a relatively new field of research for learning organizations, which appears as a trend in all the Horizon Project Reports starting with 2010, when it was part of the Visual Data Analysis field [1].

During the first International Conference on Learning Analytics and Knowledge, organized in 2011 in Canada, the concept of Learning Analytics was defined as ”the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs", as cited by Siemens and Long [2].

Friesen clarifies the two important terms in the above definition [3]:

- Data about learners: usually these data consist of the records of students' activity in LMSs, such as logging, posting and commenting messages, accessing materials, posting assignments, but also the results in previous courses or inventories of preferences.
- Optimizing and understanding learning: can be realized using a range of possible approaches to (automatically) collect data about learners from multiple sources and to interpret this collection in order to predict and improve students' future academic performance, to help those „at risk" with prompt feedback.

Learning Analytics envisages modelling learning interactions, dynamic adaptation/ personalisation of the course materials/interactions/assignments/strategies/processes based on large-scale data collection (big data), in order to improve the learning outcomes. An important amount of data is collected by LMSs, but the things become more complex when it comes to collect/analyse the interactions and communications on social media platforms which are integrated in the learning process, and also when courses are delivered not only as online or blended courses for tens of students, but as MOOCs for hundreds or thousands of distributed participants.

Siemens and Long propose the following cycle to reflect analytics in learning, starting from course level to departmental and institutional levels [2]:

- course-level: learning trails, social network analysis, discourse analysis;
- educational data-mining: predictive modelling, clustering, pattern mining;
- intelligent curriculum: the development of semantically defined curricular resources;
- adaptive content: adaptive sequence of content based on learner behaviour, recommender systems;
- adaptive learning: the adaptive learning process (social interactions, learning activity, learner support, not only content).

As Conole [4] put in her chapter „The Use of Technology in Distance Education": „Learning analytics can be used as a tool to understand learning behaviour, to provide evidence to support design of more effective learning environments, and to make effective use of social and participatory media."

Dedicated Learning Analytics modules were implemented for different LMSs: Blackboard Analytics for Learn can help in finding if student performance is dependent on the instructor's previous training; also the Brightspace LMS (formerly Desire2Learn) comes with an array of analytics capability called Insights, reporting on at risk students' differences between courses or providing metrics related to social learning [5].

As it can be seen a wide variety of data about students' participation in MOOCs are collected, starting from registration process, learning process, assessment and certification. For example, edX platform informs learners about collection of personal data and in its privacy policy a definition for personal information is provided: "…any information about yourself that you may provide to us when using the Site, such as when you sign up for a user account or enter into a transaction through the Site, which may include (but is not limited to) your name, contact information, gender, date of birth, occupation, and if you register for an ID Verified Certificate of Achievement, a driver's license or other government issued identification [6]." Additionally to this information, if a learner wish a certificate, she/he could provide payment information to third party.

In the paper several questions related to connection between learning analytics and data privacy are discussed and a model describing a possibility for realization of data privacy in MOOCs is described.

II. METHOD

The findings in this work are reached after a review of privacy policy of five existing MOOCs platforms – edX [6], Coursera [7], FutureLearn [8], CanvasNetwork [9] and Open2Study [10], starting projects and published scientific papers concerning the emerged research questions: What kind of students' private data are shared and why?, What are the privacy models provided by MOOCs?, What data is needed for Learning Analytics modules?, Are there any third party actors processing the students' data and why?, Which are the relations between Learning Analytics and Data Privacy?, Is it possible the learners to be protected? The received results are summarized via a model reflecting the data privacy in MOOCs.

III. COLLECTED DATA BY MOOCs PLATFORMS AND THIRD PARTIES

In literature and European documents four types of data are encountered: personal, private, sensitive and confidential. The term “personal data” is legally recognized and it is defined as “any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” [11]. The term “private data” is not used by juridical science, but it is often utilized in other fields of science – for example in computer science where is talking about protection of private data [12]. In this case, the term “private data” could replace0 term “personal data” in non-legal bound way and one part of researchers consider these both terms as synonyms. Anyway, private data could be a set of personal data, just these part of it that is not shared and not become public. Sensitive data is treated as special category of data in legal documents related to “racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life” [11]. Sensitive data could include personal data about any student, worker, patient, client, user [13]. Confidential data is agreed to be used only between two parties, keeping it in secret; it is not available for public use [14]. For example, the confidential data used by Cornell University is defined in University Policy 5.10 and consists of: social security number, credit card number, driver license number, bank account number, health information [15]. In privacy policies also it is talking about non-personal data as data that is indirectly connected to identification of a person through IP address, cookies and other technology form for data collection and storage [16]. In summary, the continuum of collected data for research and educational purposes is presented on Figure 1.

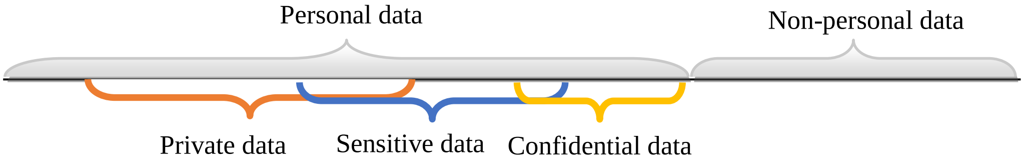

Figure 1 Continuum of collected data for research and educational purposes

In privacy policy documents of the examined five MOOCs platforms the collected data includes personal and non-personal data, but it is described and classified in different way. For example, in privacy policy of Cursera and Open2Study the collected data is explicitly divided to personal and non-personal. The rest three platforms collect non-personal data also, but it is not classified as non-personal. The term “private data” is not used in the legal policies of the five MOOCs platforms; term “personal data” is well described in the explored five MOOCs platforms; the term “sensitive data” is used only by Cursera and Open2Study. The term “confidential data” is not mentioned in any of them.

For better understanding what kind of data is collected, how and by whom this data is utilized, the five MOOCs platforms are compared according to their privacy policy documents and the findings are summarized in Table 1. Seven factors for comparison are formed: collected information by platforms, collected financial information, used information by platforms, used information by third parties, usage of cookies and other technologies, security mechanisms and used privacy policy.

Table 1 Collected and used data in MOOCs platforms according to their privacy policies

Factors/ MOOCs Platforms	**edX [6]**	**Coursera [7]**	**FutureLearn [8]**	**CanvasNetwork [9]**	**Open2Study [10]**
Collected information by platforms	Personal information about: -sign up; -participation in online courses; -registration for a paid certificate; -at sending email messages; -participation in public forums; -student learning performance; -which, when, where pages are visited; -which hyperlinks and other user interface controls are used.	A.Personal information: -at registration; - for user account update; -to purchase products or services; -to complete surveys; -to sign-up for email updates; -to participate in forums; -to send email; -to participate in online courses. B.Non-personal information: - which, when, where pages are visited; -which hyperlinks are visited; -URLs from which Cursera is visited; -log of IP address, OS and browser.	Personal information: -for access to website; -to use online courses and content; -to register or post notes, assignments or other material; -information when user report a problem; -location, gender and educational history or qualifications; -other information for delivering the best service; -results of assessments; –at purchase a paid service - credit card or other payment details; - a record of email correspondence; -data about usage of content; -IP address.	Personal information: -data for login; -messages and stored files; -email content; -survey data; - how user interacts with applications; - encountered errors by users when use platform; - device identifiers; -how often users visit the site, what pages are visited, what other sites are used for coming to the site; -IP address, operating system, browser type, domain name.	A.Personal information: -at registration; -educational qualifications ; -academic results; -banking and payment details; -tax file number; - responses to surveys; -how, when and why services are used. B. Sensitive information -language background; -citizenship; -disability; -health information. C.Non-personal information - website visit - date, time and duration, search terms, viewed pages; -IP address, type of device, browser and OS.
Collected financial information	By third party payment site	-A. By third party payment site; -B. by Coursera: credit card information	By FutureLearn: data about credit card or other payment details at purchasing a paid service	-	By Open2Study that uses EFTPOS and online technologies for payment
Usage of collected information by platforms	-to personalize the course material; -to improve learning process.	A.Personal information for: -delivering specific courses and/or services; -sending updates about online courses and	- to improve, maintain and protect the quality of web site, online courses and content; -to personalize browsing; -for announcements	- to create and maintain user account; -to provide better services; - to personalize and improve learner experience;	- for user identification and verification; -for communication; - for

		events. B.Non-Personal information for: -improvement of services; -other business purposes.	and email notifications; -for opinion gathering	- to send administrative e-mail and announcement; -to send surveys	delivering of educational services; - for improvement of web site services
Usage of collected information from third parties	-by educational institutions; -by other providers of courses in edX	- by service providers, vendors and contractors; -by university partners and other business partners; - by government authorities	- by FutureLearn partners; -by course and content providers; -other third parties	- third party service providers; - for merger, financing, acquisition, bankruptcy, dissolution, transaction, sale purposes	-by Australian government - by organizations administrating the business; - by financial institutions; -by service providers and research agencies, mailing houses, postal, freight and courier service providers, printers and distributors of direct marketing material, external business advisers
Security	-software programme for data protection through administrative, physical, and technical safeguards	-industry standard: physical, technical and administrative security measures - not sharing information with third parties, except before mentioned	-technical and organizational security measures	-secure web site - protect against unauthorized access, information use, or disclosure -any posted content using CanvasNetwork services is at user risk	-ICT security; -secure office access; -personnel security and training; -workplace policies
Usage of cookies	cookies: for collecting IP address, operating system, and browser information	cookies and/or web beacons: -to identify users; -to personalize experience; -to identify repeat visitors; -to determine the type of content and spending time	cookies: - to enhance learner experience; -for better understand how learners use the website; - cookies from third party social media websites	-cookies and web beacons: for information about user visit and searched/ viewed content - flash cookies: to store user preferences and personalize visits	cookies: -to hold anonymous session; - to personalize website visits

Privacy policy	-Own edX privacy policy; -Family Educational Rights and Privacy (FERPA) for education records	Own Cursera privacy policy	Own FutureLearn privacy policy	Own CanvasNetwork privacy policy	-Own complying with: the Privacy Act 1988, State and Territory health privacy legislation, the Spam Act 2003, the Do Not Call Register Act 2006

IV. BIG DATA AND MOOCs ANALYTICS

MOOCs platforms are a source of huge amount of data recording any activity concerning learning behavior, progress and achievements of any enrolled learner in a scale that is not recognized in traditional educational online environments. This “big data” is precisely collected and analyzed with aim to deliver personalized learning in an effective way and it is a base for predictive analytics.

O’Reilly and Veeramachaneni in their paper discuss new technology approaches and challenges for performance of MOOCs analytics when learners observe, submit, collaborate and give feedback (MOOCdb data model) [17]. They distinguish several web platforms covering different point of view in data analytics facilitating huge groups to contribute in development of analytics: MOOCviz – platform for collaborative MOOC visualizations, FeatureFactory – interactive platform for presenting prediction problems, LabelMe-Text – platform designed in support of annotating in forum posts.

Tabaa and Medouri present a learning analytics system for MOOCs: LASyM that mines “big data”, analyzes learning outcomes and assessment of learners and delivers information that could be used for design of optimized MOOCs [18]. The MOOCs learners are divided to two groups: (1) at-risk students consisting of ghosts, observers, non-completers, passive participants and (2) active participants. The aim of LASyM is to minimize at-risk MOOC students through their identification at the earliest possible stage.

Godwin-Jones is talking about the growing interest among researchers and educators to open analytical systems like edX Insights and Tin Can [19]. Their usage could improve effectiveness of course design in different subjects – for example different learning scenarios could be developed for studying computer science and art history; effectiveness of peer ranking and improvement of course assignments.

European project Multiple MOOC Aggregator (EMMA) reports development of a MOOC platform that aggregates existing European MOOC courses with possibility for learners to design their own personal learning environment [20]. EMMA platform possesses analytical application for ensuring realization of personal learning paths according to individual learning needs. The learning process is monitored, achievements are controlled and development of a competence like “learning to learn” is improved.

In the scope of other European project EDSA (European Data Science Academy) a data mining process in MOOCs on Coursera is applied with aim the learning process to be analyzed [21]. An experiment with students from Eindhoven University of Technology registered for participation in MOOCs on Coursera is performed, their behavior is analyzed and a learning model of successful and failed students is created.

Also, there are other good examples and best practices proving the importance of applying learning analytics to “big data” mining in the context of improvement of MOOCs learning scenarios. In all of these cases a huge massive of information concerning the person of any learner is utilized and personal data is not kept private. This is the reason for looking for a model improving the data privacy when a learner is involved in MOOCs.

V. CIC DATA PRIVACY MODEL IN MOOCs

The reviewed scientific literature and privacy policies of existing MOOCs platforms are a bases for development of a model describing measures for ensuring data privacy in scenarios of large scale learning. On Figure 2, the developed *CIC (Classification-Information-Choice) model* with six measures is presented coming to describe the possibilities for guarantying data privacy in MOOCs. The measures are related to:

(1) *Classification*. Accomplishment of appropriate *classification* in groups of collected personal data, *classification* in groups of collected non-personal data and *classification* of third parties according to different criteria;

(2) *Information*. Preparation of informative tools informing learners about type of collected data, mechanisms for collecting and storing, data usage and possibilities for choices;

(3) *Choice*. Performance of *choices* regarding data sharing to MOOCs platform when data are classified in different groups, *choices* concerning data giving to third parties and *choices* for movement personal and non-personal data from one classification group to other.

As it can be seen the realization of data privacy in MOOCs learning is not easy task because of involvement of multiple stakeholders and interested parties with a wide variety of interested issues applying in different context. Additionally there is a need from development of mechanisms and tools: for good data classification, for delivering complete information to learners about their collected data and for schemes in support of suitable choices in a given learning scenario.

Classification	A. Classification of collected personal data in groups according to:	1. type of data	• Personal: -private -sensitive -confidential
		2. level of importance for improvement of a learning process	
		3. level of importance for third parties	
		4. level of importance for research	
	B. Classification of non-personal data according to:	1. data type	
		2.importance for learning process	
	C. Classification of third parties in groups according to:	1. type of the third party	
		2. urgency for data utilization	
Information	D. Preparation of informative tools introducing learners to collected data at important steps in learning process concerning:	1. data type	
		2. mechanism for data gathering	
		3. data usage	
		5. possibilities for choices	
Choice	E. Choices for data sharing from different classification groups by MOOCs platform:	1. through learning process as whole	
		2. at important steps in learning process	
		3. choices concerning movement of personal data from one classification group to other	
	F. Choices for giving personal data for use from third parties according to:	1. type of the third party	
		2. purposes for usage	

Figure 2 CIC model for data privacy in large scale learning

VI. CONCLUSIONS

At this moment, researchers are intensively working on data privacy in different applicable areas in practical and theoretical aspect, including in MOOCs learning. Anyway, this topic is still in its immature form and it is a need for further development of policies, mechanisms and tools. In the existing documents the privacy of personal data is well discussed, but nothing is mentioned about the privacy of non-personal data. As it is seen a connection between personal and non-personal data exists and this connection should be taken into consideration when a data privacy is forming. The created CIC model reflects on the current situation: there are privacy policies written for a given MOOCs platform, but they are not complete; existing learning analytics tools are utilized with minimum care for data privacy; "big data" is collected and stored in unsafe digital repositories. It reveals the big picture for undertaking the measures for improvement of data privacy in MOOCs. It could be used as a recommendation tool guiding developers of MOOCs in realization of more complete data privacy.

Reference Text and Citations

[1] NMC Horizon Report Project. Higher Ed Edition, 2010-2016. Retrieved from http://www.nmc.org/publications.

[2] Siemens, G., Long, P., 2011. Penetrating the fog: Analytics in learning and education. Educause Review, 46(5), 30-32.

[3] Friesen, N., 2013. Learning Analytics: Readiness and Rewards. *Canadian Journal of Learning and Technology*, Retrieved from http://scholarworks.boisestate.edu/edtech_facpubs/95/.

[4] Conole, G., 2014. The Use of Technology in Distance Education. In Online Distance Education. Towards a Research Agenda. AU Press, 2014.

[5] Sclater, N., 2014. Analytics systems centred around the VLE/LMS. *Effective Learning Analytics*. Retrieved from http://analytics.jiscinvolve.org/wp/2014/10/08/analytics-systems-centred-around-the-vlelms-2/.

[6] MOOCs platform Edx, https://www.edx.org/.

[7] MOOCs platform Coursera, https://www.coursera.org/about/privacy.

[8] MOOCs platform FutureLearn, https://about.futurelearn.com/terms/privacy-policy/.

[9] MOOCs platform CanvasNetwork, https://www.canvaslms.com/policies/privacy.

[10] MOOCs platform Open2Study, https://www.open2study.com/legal/privacy-policy

[11] 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 281 , 23/11/1995 P. 0031 - 0050, Retrieved from http://eur-lex.europa.eu/LexUriServ/LexU.

[12] Keeping Your Private Data Secure, white paper by Symantec, Retrieved from https://www.symantec.com/content/dam/symantec/docs/white-papers/keeping-your-private-data-secure-wp-21349382.pdf.

[13] What is sensitive data? Retrieved from http://web.mit.edu/infoprotect/docs/protectingdata.pdf .

[14] UK Data Archive, Ethical/Legal/Definitions, Retrieved from http://www.data-archive.ac.uk/create-manage/consent-ethics/legal?index=4.

[15] What is Confidential Data? Retrieved from http://www.it.cornell.edu/services/guides/data_discovery/confidential_data.cfm.

[16] Out-law, Information should not be regarded as personal data if it is too burdensome to confirm its status, Council of Ministers says, 27 June 2012, Retrieved from http://www.out-law.com/en/articles/2012/june/information-should-not-be-regarded-as-personal-data-if-it-is-too-burdensome-to-confirm-its-status-council-of-ministers-says/.

[17] O'Reilly, U., Veeramachaneni, K., 2014. Technology for Mining the Big Data of MOOCs. *Research&Practice in Assessment*, vol. nine, pp. 29-37.

[18] Tabaa, Y., Medouri, A., 2013. LASyM: A Learning Analytics System for MOOCs., *International Journal of Advanced Computer Science and Applications* (IJACSA), vol. 4, No. 5, pp.113-119.

[19] Godwin-Jones, R., 2014. Emerging technologies global reach and local practice: the promise of MOOCS. *Language Learning & Technology*, vol. 18, Number 3, pp. 5–15.

[20] Brouns, F., Tammets, K. Padrón-Nápoles, C., 2014. How can the EMMA approach to learning analytics improve employability? Retrieved from http://openeducationeuropa.eu/en/article/How-can-the-EMMA-approach-to-learning-analytics-improve-employability%3F.

[21] EDSA project, Learning Analytics, Retrieved from http://edsa-project.eu/resources/learning-analytics/.