



# Extraction of Student Interaction Data from an Open edX Platform\*

Daniel Jaramillo-Morillo\*\*

Mario Solarte\*\*\*

Gustavo Ramírez-González\*\*\*\*

Received: Received: 10/10/2019 • Accepted: 16/06/2020

<https://doi.org/10.22395/rium.v20n38a5>

## Abstract

The Massive Open Online Courses (MOOC) are courses available to the general public without restrictions that are offered to hundreds or thousands of students and in recent years have been presented as a revolution in online education. They are presented as an alternative to the great demand in higher education for the characteristic of being open and massive because they allow access to education to a huge number of students. They have become an ideal environment for data collection and through the application of learning analytics techniques they have allowed a better understanding of how students learn. However, access to the data from the current open-source MOOC platforms is limited and often difficult to collect and process. This paper presents a proposal for collecting and processing the data from students' interaction with the Open edX platform through Scripts and a Collector based on Java code.

**Keywords:** data analysis; data processing; learning analytics; learning environments; learning management systems; learning platform; massive open online courses; online education; open edX.

---

\* Article derived from the project *MOOC-Maker Construction of Management Capacities of MOOCs in Higher Education* (561533-EPP-1-2015-1-ESEPPKA2-CBHE-JP) funded by the European Commission through the Erasmus+ Programme for the implementation and dissemination of the results.

\*\* Ph.D. student in Telematic Engineering, University of Cauca. Email: [dajaramillo@unicauca.edu.co](mailto:dajaramillo@unicauca.edu.co). Orcid: <https://orcid.org/0000-0001-8951-1373>

\*\*\* Ph.D. in Telematic Engineering, Associate Professor, University of Cauca. Email: [msolarte@unicauca.edu.co](mailto:msolarte@unicauca.edu.co). Orcid: <https://orcid.org/0000-0002-3600-7592>

\*\*\*\* Ph.D. in Telematic Engineering, Associate Professor, University of Cauca. Email: [gramirez@unicauca.edu.co](mailto:gramirez@unicauca.edu.co). Orcid: <https://orcid.org/0000-0002-1338-8820>

## Extracción de datos de interacción de los estudiantes en una plataforma Open Edx

### **Resumen**

Los cursos masivos abiertos en línea (MOOC por sus siglas en inglés) son cursos que están disponibles para el público general sin restricciones y que están disponibles para cientos o miles de estudiantes. Estos cursos han sido presentados como una revolución de la educación en línea. Son presentados como una alternativa a la alta demanda en la educación superior por la característica de ser abiertos y masivos y permitir la participación de una gran cantidad de estudiantes. Se han convertido en el entorno ideal para la recolección de datos y a través de la aplicación de analíticas del aprendizaje han permitido una mejor comprensión de cómo aprenden los estudiantes. Sin embargo, el acceso a los datos en las plataformas MOOC actuales de código abierto es limitado y a veces éstos son engorrosos de recolectar y procesar. Este artículo presenta una propuesta para recolectar y procesar los datos de las interacciones de los estudiantes con la plataforma Open edX a través de Scripts y un Collector basado en código Java.

**Palabras clave:** análisis de datos; procesamiento de datos; analíticas del aprendizaje; entorno de aprendizaje; sistemas de gestión del aprendizaje; plataforma de aprendizaje; cursos masivos abiertos en línea; educación en línea; open edX.

## INTRODUCTION

The growing advance and use of communication technologies have allowed the development of new educational trends based on ubiquity and networking. New forms of education have gradually become very popular, such as the Massive Open Online Course (MOOC) [1-2]. The MOOC are open and participatory courses that are offered free of cost to hundreds and thousands of students and cover topics ranging from technology to poetry. They enable a great expansion in online education. They have experienced rapid development and have received great attention from many institutions and universities. The MOOC have been presented as a new disruptive technology in the educational field [2-4].

The MOOCs have been presented as an opportunity to expand coverage of higher education and also expand access to more students. They are presented as the new path for the expansion of knowledge, university innovation, employability and the sustainable development of massive learning scenarios. This is why many universities are committed to incorporating MOOC into higher education [5-7]. This incorporation and the use of learning environments destined to the offering of MOOC has been increased thanks to the different advantages previously presented and generated from the characteristic of massiveness [8].

The popularity of MOOCs and the massive numbers of participants have made it possible to generate a large volume of data and use it for analytical purposes. Some MOOC data are the same as those obtained in a presential course such as teaching materials, demographics, student background, enrollment information, assessment results and grades [9]. Virtuality allows leaving a trace of the interaction of the students with the learning platform. This data through data analysis techniques can help to understand the behavior of students and how they learn through a virtual environment [10-12].

However, access to data from current MOOC platforms is limited and often difficult to collect and process. Aggregating learning platforms such as Coursera, Edx, MiriadaX, etc., do not easily deliver student interaction data to their associates due to different data processing policies. On the other hand, the own or private platforms, although always keeping the interaction records, do not contain tools that allow the extraction and processing of such data for analysis.

Since the first period of 2016, an instance of the learning platform “Open edX” has been implemented at the University of Cauca. An open-source platform named Selene. The first courses offered were in the modality of Small Private Online Courses (SPOC) [13-14] and Massive Private Online Course (MPOC) [15-16], variants of MOOCs that

are characterized by being limited in access (private courses) and therefore also in size but with a wider scope of participation than any conventional online course [13–17].

From the first experiences of the courses offered on the platform, the need to follow up on the students and knowing how they interact with the content arose. This is because the courses have an academic recognition and therefore the instructors in charge of the courses wanted to have a greater control than in conventional MOOCs. The follow-up and analysis of student behavior are strategies that are not incorporated into the Open edX platform. With this the following research question was posed: How to capture and process the data of the interaction of the users of an instance of Open edX?

To answer this question, the objective of this work was to design the necessary mechanisms to capture Selene's event records and process them to build a Data Set that will allow through Learning Analytics to understand the behavior of students in Selene Unicauca. The creation of several Scripts and technological development that are presented in this document is proposed. In section 2, some of the related works that were taken into account for the study are described. In section 3, the proposed prototype is presented. In section 4, the results obtained are shown. In section 5, the study presents conclusions and future work.

## 1. RELATED WORKS

In [18] and [19], the author presents a system of learning analytics with video data for a MOOC. The system captures the interactions of students with the video player (pause, repetition, forward, stop, etc.) using a Youtube API and at the same time collects information about the student's performance in terms of summative assessments. In both works, each time you press a button an abbreviation of the name of the action and the time it occurred are stored in the Google database. Interaction data and student assessment results can be displayed statistically to help tutors better understand student behavior. However, the analysis of the data is still performed by the tutors, a complicated task when the number of students is massive. Also, other types of interactions are not taken into account, such as signings on the platform and interaction with evaluations and forums.

In [11] a set of data from a course at Coursera is analyzed. Several data mining processes or techniques were used and provided some indicators in terms of usefulness, ideas, and guidance for teacher intervention in the courses to improve the quality and delivery of MOOC. In the work, they manage to obtain a classification by groups of the students. The first criterion for grouping is the type of certificate to which students are enrolled. The second criterion is the level of achievement or final grade. It is concluded that the most successful students review the contents and carry out their assessment activities in a more structured and linear way than the least successful students. It is

mentioned that data mining can help to understand student behaviors and contribute to improving course designs and the quality of education offered. However, the data is provided directly by the Coursera team. They do not perform the data extraction and processing procedure.

In [10] an exploratory study on an instance of the Open edX platform and a course offered by the University of Cuenca and the Ecuadorian Consortium for Advanced Internet (Cedia) are presented. The study analyzes the browsing behavior of students and relates it to the level of self-regulation they have and their learning style. All the results obtained from this work allow us to understand how students interact with different levels of self-regulation and different learning styles. The results suggest that MOOC should be designed to address the heterogeneity of students and for this purpose an adaptive course structure should be suggested that proposes learning activities and presents content based on the particularities of each student. However, the course was attended by 78 students of whom 24 dropped out or did not complete the course. The number of students present in the course is not sufficient to achieve a good characterization of the behavior of the students in a MOOC and does not describe the process for extracting the data from the platform.

In the previous works, it is observed how the interaction of the students with the contents can be monitored to know the behavior of the student, to improve contents, to improve the instructional design, and even to know its progress. However, none of the papers describe the process of obtaining data from the platform.

## 2. PROPOSED MECHANISM

The design of the mechanism for monitoring the learning activities of students in Massive Open Online Courses was done taking into account the structure and operation of the Open edX platform. To describe software architectures, it was decided to use the 4+1 views model. This model allows us to represent in a standard form the architecture through UML diagrams. Figure 1 shows the general architecture of the built mechanism [20].

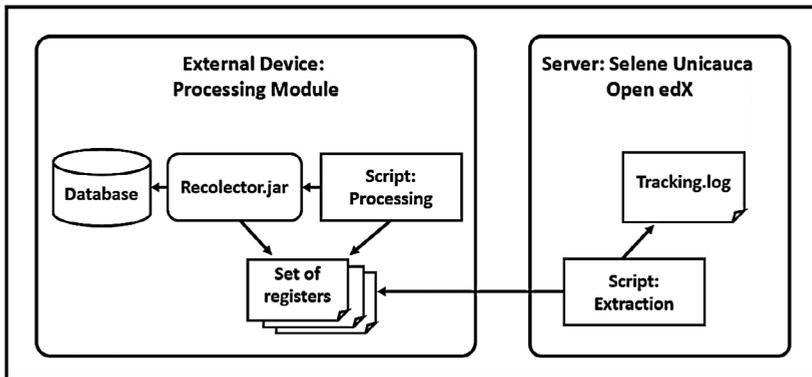


Figure 1. General architecture of the proposed mechanism.

Source: Own elaboration.

A description of each of the components of the architecture and their relationship to the other elements within the architecture is given below.

*Tracking.log*: the Selene platform is deployed on a server computer with a version of Ubuntu server 14.04 and is an instance of Open edX. A log file called tracking.log is generated on the platform. In this file are registered all the interactions of the students with the Selene learning environment, such as login, login to the courses, browsing in contents, presentation of forums, and information of the interaction with the evaluative activities. The location of the tracking.log file within the instance is: /edx/var/log/tracking/. A new log file is periodically created here depending on the amount of interaction data generated. The file is compressed and stored leaving a register of the interactions in the platform since it is released.

*Script and extraction*: scripts are a set of instructions generally stored in a text file that must be interpreted line by line in real-time for execution. The extraction script contains the commands used to make a copy of the complete folder containing the tracking.log registers of the platform and a synchronization line of this folder in an external device. In this way, all logs can be processed in real-time. Figure 2 shows the script executed every 5 minutes.

```

GNU nano 2.5.3      File: script_extraccion.sh      Modified
cp -r /edx/var/log/tracking/ /home/selene/respaldo/
rsync -avzhie "ssh -p 22" usuario@192.168.1.103:/home/recolector/Registros
/home/selene/respaldo
    
```

Figure 2. Extraction Script (Screenshot)

Source: Own elaboration.

*Script and processing:* once all the registers are available in the external device, the processing script captures the tracking.log file (a file that has the information of the students' interactions with the platform), makes a copy of it, places it in a specific address where Recolector.jar can process it and repeats this operation periodically every 5 minutes through the Cron program present in Ubuntu. Processed interaction data can be used to follow-up on students at all times. To obtain the data of past events, the script was modified to decompress one by one the registry files and to order the execution of Recolector.jar for each registry present in the synchronized folder. It was possible to create a data set with student interaction data from the moment the platform was created. Figure 3 shows the processing script which is executed every 5 minutes.

```

GNU nano 2.5.3      File: script_procesador.sh      Modified
cd /home/recolector/tracking/
rm -rf tracking.log

cp /home/recolector/Registros/tracking.log /home/recolector/tracking
java -jar Recolector.jar

echo "tracking procesado"

```

Figure 3. Log processing script (Screenshot)

Source: Own elaboration.

*Recolector.jar:* is a java-based executable. When executed, it searches for the tracking.log file in a specific location, reads it taking event b and event present in the log file, saves in a buffer the events related to the activities you want to capture, obtain the information in an orderly format and saves them in a database (MySQL). The tracking.log file is written in JSON so the collector must interpret this information properly. Figure 4 shows an example of the events registered in the tracking.log file.

*In the figure:* within the registration, the line is saved all the necessary information for monitoring activities. For example, the couple “name: play\_video” identifies that a student reproduced a video within the learning platform, including student ID information, course, date, section and subsection in the course content, etc. Thus, it is possible to capture student interaction events in terms of access, content, resources, forums, and evaluations.

```
{ "username": "Luis [REDACTED]", "event_source": "browser", "name": "play_video", "accept_language": "es-ES,es;q=0.9", "time": "2018-09-16T18:56:06.506539+00:00", "agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.106 Safari/537.36", "page": "http://selene.unicauca.edu.co/courses/course-v1:Unicauca+LaTeX_Fish+2018-II/courseware/714c09237b02479285423321dd8f326f/779d9398f56f4f6a9d931df0c3150bb7/", "host": "selene.unicauca.edu.co", "session": "ee9e0282fa511c756ebd8a2aab6cec6c", "referer": "http://selene.unicauca.edu.co/courses/course-v1:Unicauca+LaTeX_Fish+2018-II/courseware/714c09237b02479285423321dd8f326f/779d9398f56f4f6a9d931df0c3150bb7/", "context": { "user_id": 455, "org_id": "Unicauca", "course_id": "course-v1:Unicauca+LaTeX_Fish+2018-II", "path": "/event"}, "ip": "10.0.2.2", "event": "{ \"code\": \"\\1ZNPBYVCNMg\\\", \"id\": \"\\2bbf8ee0e3f948e9a992ecf87e8247a1\\\", \"currentTime\": 7.006137015735626}\", "event_type": "play_video"}
```

Figure 4. Event of the tracking.log file (Screenshot)

Source: Own elaboration.

### 3. RESULTS

From the execution of the mechanism with the registry files of Selene, the construction of a data set was achieved with more than 1,557,683 events generated from the first period of 2016 until the first period of 2018. Here is the interaction data of four courses that have been offered at the University of Cauca and have been recognized academically. Table 1 shows the number of enrollees in the Selene platform since the first period of 2016 until the first period of 2018.

Table 1. Enrolled in Selene platform

Courses	2016-I	2016-II	2017-I	2017-II	2018-I	Enrolled
Introducción al emprendimiento con Lean Startup	x	x	x	265	317	582
Comprensión de textos Argumentativos	105	110	109	103	97	524
Drones-Curso introductorio virtual FISH	x	x	133	101	106	340
Introducción a la Edición de textos científicos y literarios con LaTeX	x	102	99	104	100	405
Astronomía cotidiana	433	428	517	x	X	1378
Total Enrolled	538	640	858	573	620	3229

Source: Ownelaboration.

One of the courses that have achieved greater participation is the course “Astronomía Cotidiana” that for the first semester of 2017 managed to get 517 students enrolled, more than 10 times the number of students enrolled in a classroom course at the University of Cauca.

This course is catalogued as an MPOC for the characteristic of being private. To contribute to the understanding of student behaviors in MOOC, the student interactions of the course for the first semester of 2016 were obtained from the database. For data analysis, a CSV file was generated by querying the MySQL database. Figure 5 shows a screenshot of the analyzed CSV.

7242	Esteban	Unicauca+As	14:23:40	7/02/2017	a8a9d8f5663	pause_video	30ad398096a	f2f154481e4	sDuy0Nf2Cwg
7243	Esteban	Unicauca+As	14:23:40	7/02/2017	a8a9d8f5663	stop_video	30ad398096a	f2f154481e4	sDuy0Nf2Cwg
7244	Jhonatan	Unicauca+As	14:24:03	7/02/2017	a64eb97b2fb	pause_video	30ad398096a	f2f154481e4	ZgVJKrcD-kc
7245	Esteban	Unicauca+As	14:24:51	7/02/2017	a8a9d8f5663	pause_video	30ad398096a	f2f154481e4	ZgVJKrcD-kc
7246	Esteban	Unicauca+As	14:24:59	7/02/2017	a8a9d8f5663	pause_video	30ad398096a	f2f154481e4	ZgVJKrcD-kc
7247	Jhonatan	Unicauca+As	14:25:41	7/02/2017	a64eb97b2fb	pause_video	30ad398096a	f2f154481e4	ZgVJKrcD-kc
7252	Jhonatan	Unicauca+As	14:32:31	7/02/2017	a64eb97b2fb	pause_video	30ad398096a	f2f154481e4	ZgVJKrcD-kc

Figure 5. Events in the CSV obtained from Selene Unicauca (Screenshot)

Source: Own elaboration.

A small statistic analysis was performed with the CSV. This served as a follow-up instrument for the teacher who taught the course. Figure 6 shows the course behavior data related to the number of interactions made by students throughout the course with contents, videos, forums, and exams. Students in an MPOC with academic recognition behave around evaluative activities [21]. The students have big peaks of interaction on the dates the tests are scheduled.

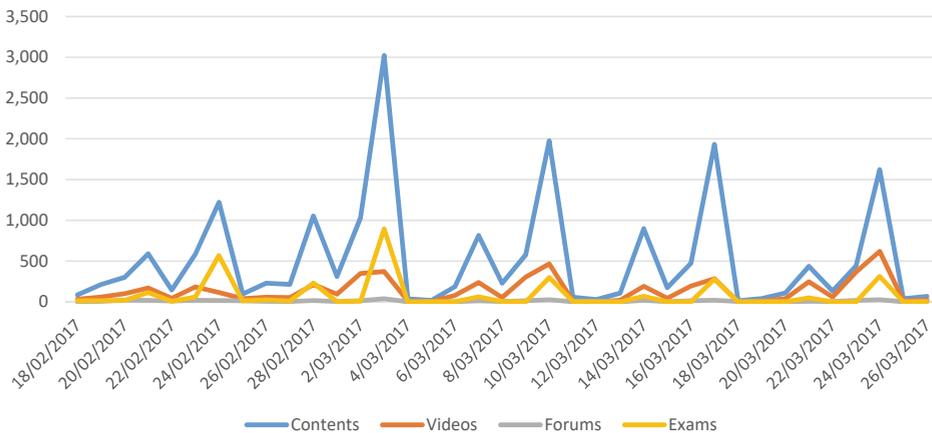


Figure 6. History of the interactions of the virtual course “Astronomía Cotidiana”

Source: Own elaboration.

#### 4. CONCLUSIONS AND FUTURE WORK

In recent years, the MOOC have positioned themselves as a new educational technology that is gradually making its way into higher education. However, there are still challenges to overcome such as offering adequate courses so that these strengthen the skills and competencies of students. An alternative that is presented is the analysis of the huge amounts of data that are obtained in the platforms that will allow us to understand how learning processes are developed in massive virtual environments.

This work managed to show the possibility and methodology used to extract the data from an instance of Open edX and to begin an analysis towards the understanding

of learning where it was possible to see the behavior of the “Astronomía Cotidiana” course in terms of student interactions. This is a good example of how students behave according to evaluative activities and leaves a question: How to make students feel motivated to perform learning activities without focusing on evaluative tasks?

It is proposed to continue working on the analysis of data collected from other courses offered through Selene that are gradually increasing and to achieve an understanding of various areas of the learning process based on variables such as student profile. Our focus in this line of research is to identify the dishonest behaviours of students in MOOC courses that have academic recognition using techniques of learning analytics.

## ACKNOWLEDGEMENTS

The authors are grateful for the support received by the project MOOC-Maker Construction of Management Capacities of MOOCs in Higher Education (561533-EPP-1-2015-1-ESEPPKA2-CBHE-JP) funded by the European Commission through the Erasmus+ Programme for the implementation and dissemination of the results set out in this article.

We would also like to thank the VRI 49694 MOOCMenTES project “Capacity Building for MOOC Management for Vocational Training, Rural Development and New Generations of Rural Students in Improving their Transit to Higher Education”, co-financed within the framework of rural partnerships by the Ministry of National Education of Colombia.

## REFERENCES

- [1] S. Downes, *Connectivism and Connective Knowledge: essays on meaning and learning networks*. Canada: National Research Council, 2012.
- [2] X. Chen, D. Barnett, and C. Stephens, “Fadorfuture: The advantages and challenges of massive open online courses (MOOCs),” Research to Practice Conference in Adult and Higher Education, Sep. 2014.
- [3] M. Gea, R. Montes, B. Rojas, and R. Bergaz, “Comunidades Activas de Aprendizaje: hacia la Formación Abierta en las Universidades,” *Vaer-Rita*, vol. 2, pp. 3, 11, Mar. 2014.
- [4] J. Kennedy, “Characteristics of Massive Open Online Courses (MOOCs): A Research Review, 2009-2012,” *Journal of Interactive Online Learning*, vol. 13, 1, pp. 1–16, 2014.
- [5] A. McAuley, B. Stewart, G. Siemens, and D. Cormier, “The MOOC Model for Digital Practice,” University of Prince Edward Island, 2010.

- 
- [6] C. M. M. García, “Diseño e implementación de cursos abiertos masivos en línea (MOOC): expectativas y consideraciones prácticas,” *RED. Revista de Educación a Distancia*, n.º 39, pp. 58-77, 2013. <http://www.redalyc.org/articulo.oa?id=54729539004>.
- [7] E. V. Cano and E. L. Meneses, “Los MOOC y la educación superior: la expansión del conocimiento,” *Profesorado. Revista de Currículum y Formación de Profesorado*, 2014. <http://www.redalyc.org/articulo.oa?id=56730662001>.
- [8] “Observatorio MOOCs UC,” 2017. [Online]. Available: <http://observatoriomooocs.sitios.ing.uc.cl/>. [Accessed: 15-Sep-2018].
- [9] U.-M. O’Reilly and K. Veeramachaneni, “Technology for Mining the Big Data of MOOCs,” *Research & Practice in Assessment*, vol. 9, pp. 29–37, 2014.
- [10] J. J. Maldonado, R. Palta, J. Vázquez, J. L. Bermeo, M. Pérez-Sanagustín, and J. Muñoz-Gama, “Exploring differences in how learners navigate in MOOCs based on self-regulated learning and learning styles: A process mining approach,” in *2016 XLII Latin American Computing Conference (CLEI)*, 2016, pp. 1–12.
- [11] P. Mukala, J. Buijs, M. Leemans, and W. Aalst, “Exploring Students’ Learning Behaviour in MOOCs using Process Mining Techniques,” Department of Mathematics and Computer Science, University of Technology, Eindhoven, The Netherlands, 2015.
- [12] H. Qu and Q. Chen, “Visual Analytics for MOOC Data,” *IEEE Computer Graphics and Applications*, vol. 35, 6, pp. 69–75, Dec-2015.
- [13] C. D. Kloos et al., “Experiences of running MOOCs and SPOCs at UC3M,” presented at the *2014 IEEE Global Engineering Education Conference (Educon)*, 2014, pp. 884–891.
- [14] A. Fox, “From MOOCs to SPOCs,” *Commun. ACM*, vol. 56, n.º 12, pp. 38–40, Dec. 2013.
- [15] W. Guo, “From SPOC to MPOC – The Effective Practice of Peking University Online Teacher Training,” in *2014 International Conference of Educational Innovation through Technology (EITT)*, 2014, pp. 258–264.
- [16] A. M. Mutawa, “It is time to MOOC and SPOC in the Gulf Region,” *Educ. Inf. Technol*, pp. 1–21, 2016.
- [17] J. Cabero, C. Llorente, and A. Vázquez, “MOOC’s typologies: Design and educational implications,” *Profesorado*, vol. 18, pp. 13–26, 2014.
- [18] K. Chorianopoulos and M. Giannakos, “Merging learner performance with browsing behavior in video lectures,” *WAVE 2013 workshop LAK’13*, 2013.
- [19] C. D. Kloos, P. Muñoz-Merino, and M. Muñoz-Organero, “Extendiendo Google Course Builder mediante Proyectos Realistas en un Curso de Master,” Universidad Carlos III de Madrid, 2014.

- [20] D. Jaramillo-Morillo, M. S. Sarasty, G. R. González, and M. Pérez-Sanagustín, “Follow-Up of Learning Activities in Open edX: A Case Study at the University of Cauca,” in *Digital Education: Out to the World and Back to the Campus*, 2017, pp. 217–222.
- [21] M. Solarte, G. A. Ramírez, and D. A. Jaramillo, “Access habits and assessment results in massive online courses with academic value.” *Revista Ingeniería e Innovación*, vol. 5, n.º 1, May 2017.