Hayashi, Y., et al. (Eds.) (2017). Workshop Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

Integration of Learning Analytics Research and Production Systems While Protecting Privacy

Brendan FLANAGAN[a*], Hiroaki OGATA[a]

[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*

*flanagan.brendanjohn.4n@kyoto-u.ac.jp

Abstract: Learning analytics researchers often face problems when dealing with data that contains personally identifying information, and the protection of stakeholder privacy in analysis systems. As learning management systems become more important within education institutions, these systems are being subject to increasingly stringent standards to protect user privacy. This however has the potential to hinder learning analytics research because data collected in production cannot simply be transferred as is to research systems for real-time analysis. In this paper, we propose a system design that provides an interface between integrated production and research systems to allow user authentication, information, and learning analytics results to be seamlessly transferred between systems. The interface provides a level of anonymity to allow a greater degree of research freedom when analyzing data without exposing private data directly through research systems.

Keywords: Learning analytics, anonymized data analysis, seamless learning

1. Introduction

In recent years, Learning Management Systems (LMS) have become an integral part of higher education. As these services are becoming increasingly important to education, LMS are being managed as production environments with stringent security and processes to safeguard the integrity of the system. While data from LMS and other VLE (virtual learning environments) are essential to learning analytics research, a particular concern is the protection of data and privacy throughout the analytics workflow (International Organization for Standardization, 2016). On one hand, researchers must ensure that the privacy of key stakeholders, such as: students, teachers, and administrators are protected. On the other hand, the protection of data privacy can sometimes limit access to data, which can hinder learning analytics research.

This problem also raises issues when production and research learning environment systems are integrated during the development of new learning analytics research ideas, and performing experiments to evaluate their effectiveness in the field. Ideally, research systems would pre-emptively protect data and privacy by only handling anonymized data that has been stripped of information that can identify a person. However, this solution also has limitations as it can negatively impact personalized results, such as: a student comparing their personal progress in a course with that of the whole student cohort. There are also possible secondary uses of data collected by these systems that should be investigated, such as: the use of real data in learning analytics and data science education, community based learning analytics where data is available to stakeholders to freely perform their own analysis, and facilitating 'data takeout' where the stakeholder can export their personal data and transfer it to another system.

Traditionally, there has been little distinction made between the different roles that systems perform, with LMS and learning analytics systems inhabiting the same environment without abstraction. However, as LMS and learning analytics research mature, systems are becoming increasingly modular with personal data being stored in numerous locations, and anonymity by design will play an increasingly important role in the protection of personal data in integrated systems.

In this paper, we propose the design of integrated production and research learning systems that address the protection of stakeholder privacy, while trying to minimize the limitations of anonymized data analysis in research systems. We are currently in the process of developing and

testing parts of a system based on this design with an anticipated small scale soft launch of the research systems from October 2017. The design presented in this paper is limited to the current requirements at hand, and does not try to address other possible requirements, such as: incorporating single-sign-on authentication which is left for open discussion. A long-term goal of this research is to implement the proposed system across various educational institutes ranging from K12 schools through to high education.

2. Overview of the Proposed Integration of Production and Research Systems

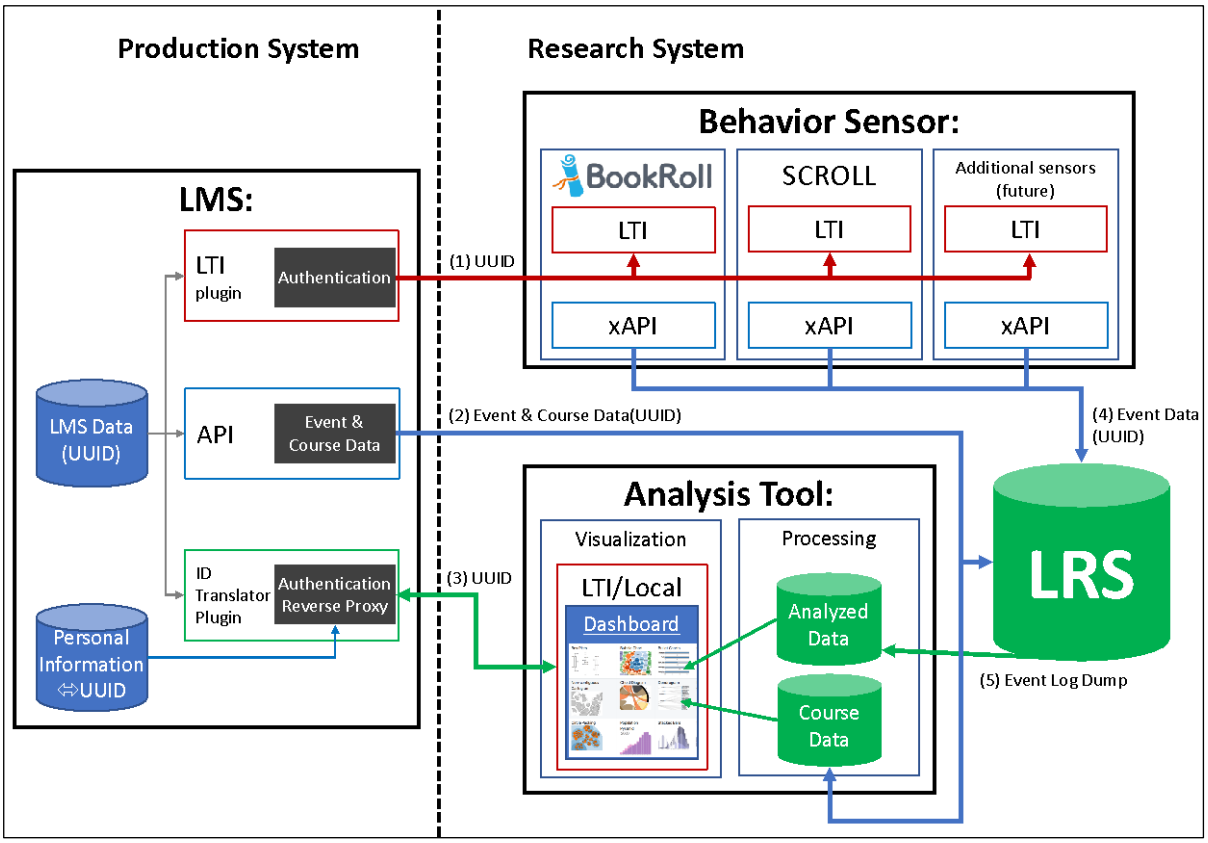

Figure 1. Overview of proposed design to integrate production and research based systems.

2.1. Learning Management System (LMS)

In recent years, several interfaces have been proposed to allow the seamless and secure integration of external tools to augment existing LMS experiences. Some of these interfaces have been proprietary and thus limited the tools that can be integrated. IMS Global Learning Consortium (2016) published the Learning Tools Interoperability (LTI) standard for defining the process of connecting two systems, and how users will transition across these systems without having to authenticate once again with the destination system. During the LTI transition process, information about the user and the context in which the external tool was launched can be transferred from the source system to the target system. In many cases, personal information is usually transferred to the target system in this process. However, this can pose a problem when production systems are integrated with research systems. Personal information is usually handled in production systems that have been designed and secured to avoid breaches of user privacy. In contrast to this, research systems are generally not concerned with the design and security aspects required to ensure user privacy. This is influenced by various factors, including: the purpose of the system, time and funding constraints, and the fact that the design and management is usually carried out by a wide range of users from highly experienced

professors to students who are just starting their first research. Because of these reasons, it is important to consider how user privacy can be protected when integrating production and research systems.

2.1.1. Anonymized Id Management

We propose that the information that is transferred when connecting external tools should be limited to attributes that cannot directly be used to identify a user as a particular person. Most modern LMS utilize an internal universal unique identifier (UUID) to which personal information, such as: real name, student/teacher id, and email address are attributed. As shown in Figure 1, we propose that (1) UUID should be the only user identification information that is transferred to research systems. The relation between the LMS's internal UUID and personal information is only available within the production system and therefore reduces the risk of a user privacy breach. External tools will then attribute learner events with the LMS's internal UUID that is sent during the LTI launch process, therefore anonymizing (4) Event data collected in the research system side LRS (Learning Record Store).

Anonymized (2) Course and event data using the LMS internal UUIDs in place of personal information will also be exported from the LMS to an analysis tool and LRS. A simple plugin within the LMS is being developed to translate the UUIDs displayed in research system analysis results into the real name, id, or email address of students and teachers. The plugin will act as a LTI Tool consumer reverse proxy, which involves both authentication using (3) UUID with the LTI Tool provider, and translating UUIDs by retrieving the contents from the provider instead of the user directly transitioning to the external tool. This ensures that the students and teachers will be able to meaningfully interpret research system analysis. This is particularly important for research into predicting at risk students as anonymized results would be difficult to use for intervention support.

2.2. Behavior Sensors

The actions in tasks that learners take during the course of their studies that occur outside the LMS need to be captured by *behavior sensors*. These tasks can take place in both formal and informal learning situations in seamless learning environments (Uosaki et al. 2013), and therefore it is important to collect data on the events that occur in both of these environments. We currently plan to implement the addition of two behavior sensor systems: a digital learning material reader called BookRoll, and an informal language learning tool called SCROLL (Ogata et al. 2011). The design of the system allows additional behavior sensors to be integrated into the proposed system. Currently the planned behavior sensors are proprietary independent systems and do not support open interoperability with other systems. We are currently developing standardized interfaces based on: LTI for seamless authentication transition from existing production LMS by anonymized (1) UUID, and xAPI (Advanced Distributed Learning, 2016) which is an open source statement API for outputting anonymized (4) Event data to a centralized independent Learning Record Store (LRS). As the main purpose of the data collected by *behavior sensors* is for research analysis, all users of the systems will be giving the option to opt-out on initial authentication if they do not consent to participation and will not have their actions logged.

2.2.1. *Digital Learning Material Reader*

Figure 2. A screenshot of the BookRoll digital learning material reader that will be deployed.

Digitized learning materials are a core part of modern formal education, making it an increasingly important data collection source in learning analytics. The reading behavior of students has previously been used to visualize class preparation and review patterns by Ogata et al. (2017). The digital learning material reader can be used to not only log the actions of students reading reference materials, such as textbooks, but also to distribute lecture slides, etc. Real-time analysis of students reading lecture slides can be visualized to inform a teacher that they need to slow down if too many students are reading previous slides of the current slide that is being explained. Conversely, the teacher may need to speed up if too many students are reading ahead of the current slide. Additionally, the reading logs could be analyzed to evaluate and find sections of learning materials that need to be revised. In the proposed system, we plan to deploy the BookRoll digital learning material reading system. As show in Figure 2, there are features to highlight sections of reading materials in yellow to indicate sections that were not understood, or red for important sections. Memos can also be created at the page level or with a marker to attach it to a specific section of the page. Users can also bookmark pages or use the full text search function to find the information they are looking for in later revision. Currently, learning material content can be uploaded to BookRoll in PDF format, and it supports a wide range of devices as it can be accessed through a standard web browser.

Initially, user behavior was logged in a local database and required that analysis be performed by either connecting directly, or exporting data from the database. In the proposed system, user behavior events will be sent by an xAPI interface and collected in a central independent LRS. The frequency and amount at which events will be sent will be configurable to enable either cost effective digest logging were a large number of events are sent in one request, or high frequency logging that is required for real-time learning analytics visualization.

2.2.2. *Informal Language Learning Tool*

In addition to collecting data on user behavior in formal learning situations, we also plan to deploy the SCROLL ubiquitous learning log system that was reported in Ogata et al. (2011) to collect data on user behavior in informal learning environments. SCROLL can be used to support the sharing and reuse of ubiquitous learning logs that are collected in the context of language learning. The addition of behavior sensors that capture event information outside traditional formal classroom contexts enables the support of research into seamless learning analytics of language learners. As the proposed system will collect data from both formal and informal learning environments, this will enable linking of

knowledge learnt in either context in addition to information from the LMS, and could be analyzed to predict and extract behaviors of overachieving and underachieving language learners.

Additional integration of specialized language learning tools, such as: testing and exercise systems for the four major skills: listening, speaking, reading, and writing, into the proposed system would provide further opportunities to analyze in detail the behavior of language learners, however at the time of writing this is beyond the scope of this paper and will be addressed in future work.

2.3. Learning Record Store (LRS)

The LRS is an integral part of the proposed system as it will be a central independent point to collect all event data from both the production LMS system and behavior sensors which are still in the research phase of the development cycle. While we have chosen to adopt xAPI as the mode of transporting events data from other systems to the LRS, this is not a strict limitation. We have decided to deploy the latest version of Apereo Foundation's OpenLRS (Apereo Foundation, 2017), which has the ability to support the storing and querying of event data from both xAPI and Global Learning Consortium's Caliper Analytics API (2015). Data from both interfaces are stored in a unified format within the LRS, which will aid data analysis as researchers will not have to spend as much time extracting, transforming, and loading data (ETL). The collection of data in an LRS also reduces information silos were data is only stored locally in a number of different modular systems, and has the potential to increase the availability of data for analysis. In the proposed system, we plan to automate the ETL process by taking incremental (5) Event log dumps from the LRS database as seen in Figure 1, and sending it to the Learning Analytics Tool for automated processing.

2.4. Learning Analytics Tool

The Learning Analytic Tool will act as a dashboard portal system to display actionable results and outcomes of learning analytics in the form of visualizations. The portal is intended to serve a number of different stakeholders, from students comparing their individual progress against that of their anonymous peers, teachers checking the overall progress of the classes under their care, to administrators surveying the effectiveness of education they are offering in their institution. It is proposed that students and teachers will access the portal via a plugin within an LMS that will provide both authentication of the user and also translate the UUIDs that are displayed in the portal into their corresponding real identities depending on their role in the LMS. Teachers who are in charge of class will be able to view all the student identities of students within that specific class. However, students will only be able to view their own identity, and the identities of their peers will remain anonymous in the results of the analysis. Administrators will login into the portal through a local authentication system, and the visualizations will only contain anonymized results that protect the identities of individuals.

This tool will be split into two main parts. The first part is a processing system that will analyze raw (5) Event log dumps from the LRS along with (2) Event and course data from the LMS. This process will extract and calculate relevant metrics for actionable results and outcomes and store these in a local database for analyzed data. The second part is a visualization system platform which will host customizable visualizations of the analyzed data. The UUIDs that are displayed in the portal will be marked up with tags to enable quick and effective parsing and translation to the real identities by a plugin within the production LMS system.

3. Conclusion

In this paper, we propose the design of integrated production and research learning analytics systems where personal information is only stored in the production system. We address issues on user privacy by proposing the use of an LMS's internal UUIDs to anonymously collect and analyze learner behavior while using research systems. The visualizations of outcomes and actionable results from the research systems can then be viewed via a reverse proxy plugin that resides within the production

LMS system, and translates the anonymous UUIDs into the real identities based on the users' role within the LMS system.

An advantage of the proposed system is that as the data collected by the system does not contain information that can directly identify students, it allows the data to be openly analyzed within the connected research systems. In the future, we plan to allow students of courses, such as: learning analytics and data mining, to analyze the real data collected by the proposed system. We expect this will help in the development of education of these fields, and encourage students to pursue further research and analysis of their own learning behavior.

In future work, we will complete the implementation of the system and evaluate its effectiveness in meeting the needs of students, faculty staff, and researchers.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 16H06304.

References

Advanced Distributed Learning. (2016). Experience API (xAPI) Specification. Retrieved from http://github.com/adlnet/xAPI-Spec/

Apereo Foundation. (2017). Apereo Learning Analytics Initiative: Open LRS. Retrieved from the website of Apereo Foundation http://www.apereo.org/projects/openlrs

IMS Global Learning Consortium. (2015). Caliper analytics. Retrieved from the website of IMS Global Learning Consortium http://www.imsglobal.org/activity/caliper

IMS Global Learning Consortium. (2016). Learning Tools Interoperability (LTI). Retrieved from the website of IMS Global Learning Consortium http://www.imsglobal.org/activity/learning-tools-interoperability

International Organization for Standardization. (2016). Information technology for learning, education and training -- Learning analytics interoperability — Part 1: Reference model (ISO/IEC TR 20748-1:2016). Retrieved from http://www.iso.org/standard/68976.html

Ogata, H., Li, M., Hou, B., Uosaki, N., El-Bishouty, M. M., & Yano, Y. (2011). SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. Research & Practice in Technology Enhanced Learning, 6(2), 69-82.

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., ... & Hirokawa, S. (2017). Learning Analytics for E-Book-Based Educational Big Data in Higher Education. In Smart Sensors at the IoT Frontier (pp. 327-350). Springer International Publishing.

Uosaki, N., Ogata, H., Li, M., Hou, B., & Mouri, K. (2013). Guidelines on Implementing Successful Seamless Learning Environments. International Journal of Interactive Mobile Technologies, 7(2).