Scalable Higher Education Learning Analytics Architecture through Data Integration

Jeanette Samuelsen[1][0000-0002-8084-0258]

[1] University of Bergen, Bergen, Norway
jeanette.samuelsen@uib.no

Abstract. Effectively implementing LA at an institutional level is far from trivial, as such a solution needs to be scalable. In this project I aim to build a scalable learning analytics architecture. The main research focus will be on semantic interoperability, for integration of multiple educational data sources, but the architecture will also have components for data analysis and reporting insights to stakeholders. As part of this project we have conducted a systematic review, to understand state-of-the art research and practice regarding multiple data source usage and combination in learning analytics. The initial work with the architecture has also resulted in a first conceptual model. For developing the architecture, semantic web technologies will be employed, to handle aspects such as shared data meaning. Once developed, we will conduct studies that use the architecture to address real-world challenges in higher education.

Keywords: Learning analytics, Interoperability, Semantic web, Scalability

1 Introduction

Learning Analytics (LA) includes collecting, computationally analyzing and reporting data to stakeholders; to gain insights, and enable decision making and interventions related to questions about learning and learning environments [1]. Effectively implementing LA at an institutional level is far from trivial, as such a solution needs to be scalable. Factors affecting scalability include the technical solution, but also factors such as organizational hierarchy [25], management structures [4], policy and regulations [6]. Making LA scalable includes collecting and integrating data from multiple data sources, which can be stored in different formats, and have varying levels of structure. The merge of data from many different sources can lead to more useful analysis, since many LA techniques require large scale and possibly diverse data [3].

Data integration is related to interoperability. Interoperability involves technical, semantic, legal and organizational levels. The semantic level is about ensuring that data format and meaning is preserved and understood. The technical level includes services for data exchange and data integration. The legal level is about aspects such as enabling collaboration despite of different organizational policies. Organizational interoperability includes aligning processes for common organizational goals and addressing user expectations and requirements [19].

The number of data sources used in combination for data analysis tend to be limited. Existing LA projects addressing data integration needs tend to put the emphasis on the technical level of interoperability [14]. The new EU general data protection regulation [6] will to a large degree address legal and organizational concerns. Semantic interoperability, enabling shared data meaning, is typically less emphasized, even though it will enable more effective merge of data.

2 Goals and Research Question

The main goal of this project is to build a scalable LA architecture for higher education, with research emphasis on semantic interoperability. However, the architecture we will develop also has components for data analysis and reporting of insights to stakeholders. A secondary goal, providing insights into the real-world application of the architecture, will be to address one or more challenges in a higher education institution, through architecture usage. Initially, the challenges will focus on student success.

To achieve the goals, the following overarching research question has been formulated:

- How can a scalable learning analytics architecture be built and applied to address challenges in higher education?

3 (Abbreviated) State of the Field

Different studies have used or combined multiple data sources for LA. Data from a Learning Management System and Student Information System have been combined to detect students who struggle academically [13]. Researchers joined four data sets containing student data, originating from two separate tools, to build a model to predict low academic performance [9]. Behavior data from an LMS and university course database have been joined, to cluster blended learning courses [22]. A commonality among these studies is that they all combine a limited number of data sources, in similar formats.

Merging data available in different formats is more challenging than combining similar data. In addition to common operations such as data access, data cleaning, and data filtering, the data to be combined also needs to be transformed into a common format. For this purpose the organization JISC has developed an architecture that includes plugins to transform educational data to xAPI statements, meaning they can be combined in a common store (learning record warehouse) [14]. xAPI, together with IMS Caliper, are educational data specifications, both enabling standardization [7].

While the approach taken by JISC will enable technical interoperability, there is less of an emphasis on enabling shared meaning between data. Using semantic technologies, such as the RDF data model and ontologies, it is possible to add descriptions and meaning to data coming from various sources, and to combine, support and reuse different specifications/data models. Ontologies also enable inference (given some stated fact, we can state new and related facts) [10]. In addition, the use of ontologies is different from the more traditional approach of data warehousing [15], since alignment (mapping

between concepts in different systems) can be optimized through ontology reuse, rather than ad-hoc for each specific use case.

4 Research Design

As an overarching research methodology for this research project, the design science research framework will be used [13]. This methodology acknowledges a technical artifact, including development and evaluation, as a research contribution. Novel parts of such an artifact can also be viewed as contributions.

4.1 Systematic Literature Review

To identify important foundations for the work on the architecture, and to make it clear where the proposed research fits into the body of knowledge, we have recently conducted a systematic review. The systematic review follows specific guidelines [16]. Our review research questions include:

- How and to what extent are different types of data being used/combined for learning analytics research in higher education?
- What types of data are being used for learning analytics in higher education?

4.2 System Architecture

Informed by the results of the systematic literature review, we are developing a LA architecture. Of special importance is that the solution should scale, emphasizing semantic interoperability, to remove barriers for data exchange.

4.3 Studies Using the Architecture

Once the system architecture has been developed, and we have combined educational data sets, we will use the architecture to address one or more challenges in higher education, through computational data analysis and reporting. Challenge selection will be informed by stakeholder needs for specific insights about learners and their environments. Initially, the focus will be on LA related to student success (e.g. through providing relevant dashboards to the students).

5 Current Status and Results

5.1 Systematic Literature Review

The following search string was formulated to search relevant databases (ACM, IEEE Xplore, SpringerLink, Science Direct, Wiley and AISEL), conference proceedings (Learning Analytics and Knowledge, Learning at Scale and International conference on

Educational Data Mining) and journals (Journal of Learning Analytics and Journal of Educational Data Mining) for documents:

("multiple data sources" OR multimodal OR "multi-modal" OR "multiple data sets" OR "multiple datasets") AND ("learning analytics" OR "educational data mining") AND "higher education"

A number of inclusion and exclusion criteria were formulated, to ensure that reviewed papers would fit the research questions and would provide foundations for our development of a LA architecture. One such criteria was that the reviewed papers had to use or combine more than one data source.

The search originally returned 71 papers, but following specified inclusion and exclusion criteria, we were left with 14 papers for inclusion in the review. This process was documented using a PRISMA flow diagram [18].

In our results, we have observed that five of the reviewed papers analyze data in different formats that originate from different sources without a common format, thus these data are not integrated, but analyzed separately. Nine of the studies merge and analyze data from different data sources that are already in the same format. A more extensive list is given in Table 1. With regards to number of data sources used/combined, we found that nine of the fourteen papers use or combine only two data sources.

Table 1. Multiple dataset use - preliminary observations

Observation	Freq.	Papers
Merge data that are already in the same format from different data sources	9	[5, 8, 9, 12, 13, 22, 23, 24, 29]
Analyze data in different formats from different sources without a common format	5	[20, 21, 27, 28, 29]
Support educational data models/controlled vocabularies	2	[5, 21]

Having conducted the review, we are now finishing up the resulting paper and plan to send it for review before the end of the 2018.

5.2 System Architecture and Usage

The initial work with the architecture has resulted in a first conceptual model.

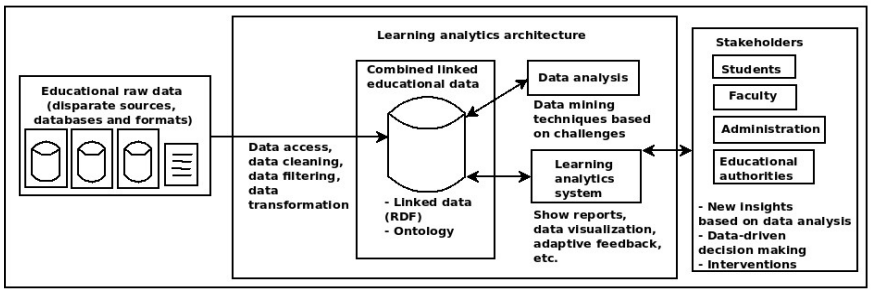

Fig. 1. High-level LA architecture (functionality) and its environment

As seen in Fig. 1, the architecture shall be able to combine a number of data sets with a multitude of formats, supported by semantic technologies. For this purpose we need to develop tools to obtain, clean and transform the relevant data. An ontology will be constructed to enable alignment between concepts in different systems. Such an ontology can build on pre-existing educational data models, but will likely need to be further extended with new concepts to address gaps in the existing specifications. To persistently store and combine educational data from different sources, a RDF database will be used. With this approach we can support data expressed in compliance with both xAPI and Caliper specifications, but also numerous other data models. In this sense the architecture goes beyond the approach taken by organizations such as JISC, and it is more streamlined than data warehousing.

We have just recently obtained educational datasets from a Norwegian higher education institution, thus development of the architecture will soon commence. Studies that use the architecture will follow.

References

1. 1st International Conference on Learning Analytics and Knowledge 2011 (February 2011), https://tekri.athabascau.ca/analytics/
2. Chang, C.J., Chang, M.H., Liu, C C., Chiu, B.C., Fan Chiang, S.H., Wen, C.T., ..., Chai, C.S.: An analysis of collaborative problem-solving activities mediated by individual-based and collaborative computer simulations. Journal of Computer Assisted Learning **33**(6), 649-662 (2017).
3. Cooper, A., Hoel, T.: Data Sharing Requirements and Roadmap (2015).
4. Dawson, S., Poquet, O., Colvin, C., Rogers, T., Pardo, A., Gasevic, D.: Rethinking learning analytics adoption through complexity leadership theory. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 236-244). ACM (2018).
5. Di Mitri, D., Scheffel, M., Drachsler, H., Börner, D., Ternier, S., Specht, M.: Learning pulse: a machine learning approach for predicting performance in self-regulated learning using multimodal data (pp. 188–197). ACM Press (2017).
6. EU GDPR (2016). http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN.
7. Griffiths, D., Hoel, T.: Comparing xAPI and Caliper. LACE Review 7 (2016).
8. Gray, G., McGuinness, C., Owende, P., Hofmann, M.: Learning Factor Models of Students at Risk of Failing in the Early Stage of Tertiary Education. Journal of Learning Analytics, **3**(2), 330–372 (2016). https://doi.org/10.18608/jla.2016.32.20
9. Guarín, C.E.L., Guzmán, E.L., González, F.A.: A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje **10**(3), 119-125 (2015).
10. Hebeler, J., Fisher, M., Blace, R., Perez-Lopez, A.: Semantic web programming. Wiley (2009).
11. Hevner, A., March, S., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly **28**(1), 75–105 (2004).
12. Hutt, S., Hardey, J., Bixler, R., Stewart, A., Risko, E., D'Mello, S.: Gaze-based Detection of Mind Wandering during Lecture Viewing (2017).

13. Jayaprakash, S., Moody, E., Lauría, E., Regan, J., Baron, J.; Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. Journal of Learning Analytics, **1**(1), 6–47 (2014). https://doi.org/10.18608/jla.2014.11.3
14. JISC (2018). https://docs.analytics.alpha.jisc.ac.uk/docs/learning-records-warehouse/Technical-Overview:--Integration-Overview
15. Kimball, R., Ross, M.: The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons (2011).
16. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01. School of Computer Science and Mathematics, Keele University (2007).
17. Liu, M., Kang, J., Zou, W., Lee, H., Pan, Z., Corliss, S.: Using Data to Understand How to Better Design Adaptive Learning. Technology, Knowledge and Learning **22**(3), 271-298 (2017).
18. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: The PRISMA group. Preferred reporting items for systematic review and meta-analysis: the PRISMA statement. PLoS Medicine 6(7), e1000097 (2009).
19. New European Interoperability Framework (2017). https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf
20. Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., Castells, J.: The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors (pp. 360–364). ACM Press (2018). https://doi.org/10.1145/3170358.3170406
21. Pardos, Z.A., Whyte, A., Kao, K.: moocRP: Enabling Open Learning Analytics with an Open Source Platform for Data Distribution, Analysis, and Visualization. Technology, Knowledge and Learning **21**(1), 75-98 (2016).
22. Park, Y., Yu, J.H., Jo, I.H.: Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. Internet and Higher Education **29**, 1-11 (2016).
23. Raca, M., Tormey, R., & Dillenbourg, P.: Sleepers' Lag: Study on Motion and Attention. Journal of Learning Analytics, **3**(2), 239–260 (2016). https://doi.org/10.18608/jla.2016.32.12
24. Rodríguez-Triana, M., Prieto, L., Martínez-Monés, A., Asensio-Pérez, J., Dimitriadis, Y. The teacher in the loop: customizing multimodal learning analytics for blended learning (pp. 417–426). ACM Press (2018). https://doi.org/10.1145/3170358.3170364
25. Shum, S., McKay, T.: Architecting for Learning Analytics: Innovating for Sustainable Impact (2018).
26. Sclater, N., Peasgood, A., Mullan, J.: Learning analytics in higher education: A review of UK and international practice (2016).
27. Thompson, K., Kennedy-Clark, S., Wheeler, P., Kelly, N.: Discovering indicators of successful collaboration using tense: Automated extraction of patterns in discourse. British Journal of Educational Technology **45**(3), 461-470 (2014).
28. Wang, Y., Paquette, L., Baker, R.: A Longitudinal Study on Learner Career Advancement in MOOCs. Journal of Learning Analytics, **1**(3), 203–206 (2014). https://doi.org/10.18608/jla.2014.13.23
29. Zheng, M., Bender, D., Nadershahi, N.: Faculty professional development in emergent pedagogies for instructional innovation in dental education. European Journal of Dental Education **21**(2), 67-78 (2017).