Chen, W. et al. (Eds.) (2016). Workshop Proceedings of the 24th International Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education

Time Well Spent: A Data Warehouse Solution for Temporal Analyses of Educational Data

Kolja P. Debus[a], Michael D. Kickmeier-Rust[a]*, Dietrich Albert[a]
[a] *Graz University of Technology, Austria*
* *michael.kickmeier-rust@tugraz.at*

Abstract: Education is an important area for data analytics; recent buzz words are Learning Analytics and Educational Data Mining. This paper present the work accomplished in the context of the Lea's Box project. Two main work strands are presented; first a technical solution for an educational data mining and learning analytics data warehousing concept that completely rests on open source software will be presented, together with its future development plans to make it accessible as a service for various data input points via xAPI. After a reflection on the obstacles faced by research that collects data from real-live educational environments, together with lessons learned by the project and possible solutions for future projects, temporal analyses of single tasks carried out by students in US middle school are presented.

Keywords: data warehouse; learning analytics; visualization; xAPI; MySQL; interconnectivity

1. Introduction

Lea's Box 1 (www.leas-box.eu) is a European research and innovation project that brings together two industry partners from Turkey and the Czech Republic with the University of Birmingham, UK and the Technical University of Graz. It is, in part, a continuation of the successful work of the Next Generation Teaching, Education and Learning for Life (NEXT-TELL) project that ran from 2011 until 2014. Lea's Box was designed to undertake research and development of tools that serve practitioners both in formal as well as informal sectors in the educational world in real-live, i.e. technology sparse, environments with little or non-existing knowledge of software handling. In the focus of the project is using educationally relevant data to optimize teaching by providing teachers with deeper insights into the learning processes of individual students. In contrast to typical (data rich) Learning Analytics scenarios (e.g., in eLearning courses), the project is focusing on the (data lean) settings of typical school education. Thus, the project envisions developing simple tools that allow teachers to bring all the available bits of data from various sources together, to analyze and visualize them in a suitable way. One key part, which is presented here, is having a broad data model in place that allows working with incomplete and extremely diverse data and, perhaps more importantly, that is able to cover all the various needs of teachers – independent of school type, age level, proficiency, or school system.

In this paper we present a data analysis warehouse solution and a complex data model that is to be understood in this context of need for stable, simple and meaningful software solutions. Its main purpose is to serve as a central hub of diverse sources of information that needs to be transformed into insight about learners, used by themselves as by teachers or administrators. In this way it is comparable, in nature, to the integrated designs (Masud and Huang, 2012; Rayon, Guenaga, and Nuñez, 2012) . Moreover, as the xAPI (https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md OR http://experienceapi.com/) is used as the connection to data sources, it holds the capacity for future tools to put data into and obtain it from the database, as needs arise. From a psychological/educational viewpoint, it serves as a transmission belt to generate a sound and high-speed basis for formative assessment about competencies, tasks and learning assignments, carried out by individual as well as groups of students. It should, in the near future, broaden the capability of tools like *myClass*[3], the Open Learner Model (Bull et al., 2016) and other existing educational software. Additionally, it can enhance the availability of visualization services such as.

[3] Accessible at www.leas-box.eu

Regarding the focus of the analysis of this paper, temporality in quantitative measures of student's actions, a perspective on the individual on a micro level (referring to the diagnosis within the fine facets of behaviors in learning objects, exercises, test items, etc.) seems to be missing in the current literature. Although there have been attempts to remedy the situation and increase the awareness for the benefit and necessity of a conceptual inclusion of a timely perspective that informs about order and the arising narrative (cf. Reimann, 2009) or a proof of concept in a qualitative empirical way, the common analysis takes the form of a longitudinal analysis on the meso (i.e., learning objects level) or macro (i.e., course) level. Such grand-scale analysis (Jo, Kim, and Yoon, 2014) is undoubtedly with merits, but more informative for administrators or fit for large scale courses as in universities or vocational schools in China. The insights to be gained by such a micro level analysis and the usage of them in classroom situations or for individual learner's advance of their study is thus presented in the final section of this paper.

2. Database

2.1 Concept

The database described in this paper was designed and deployed as part of the multi-partner international research and development project Learning Analytics Toolbox (referred to from here on as Lea's Box). The project:
"[...] aims at a competence-centered, multi-source learning analytics methodology based on the foundations of sound psycho-pedagogical models, intelligent model-based reasoning services, innovative visualization techniques, tailored to the very concrete demands and requirements of teachers and learners" (unpublished description of work for the project), see also (Kickmeier-Rust, Bull, and Albert, 2016). Fig. 1 displays the overall concept of the project. The database of this paper is located in the lower right, to provide central data filtering, data aggregation and data storage capability.

The design had three focal points: (1) simplicity at first, with simple and basal data flow destination points within the database architecture, (2) high capacity for different sources, that would create a diverse and sometimes unforeseeable influx of data, (3) transferability, since the project needs to publish all code as open source after the wrap-up, so the ease of putting the database to work somewhere else was deemed of high importance.

As MySQL was used at other parts of the project, it was deemed sufficient for the task at hand. This may be unconventional, since other database management systems were specifically designed to match the requisites of a data warehouse/data mining task, as MonetDB (Idreos et al., 2012), MongoDB (Rayon, Guenaga, and Nuñez, 2012) or cloud architectures (e.g., Sonwalkar, 2013).

Nonetheless, MySQL has the huge advantage of a widespread use, low threshold due to non-existent costs accompanied by an exhaustive documentation and quick transferability to other database management systems.

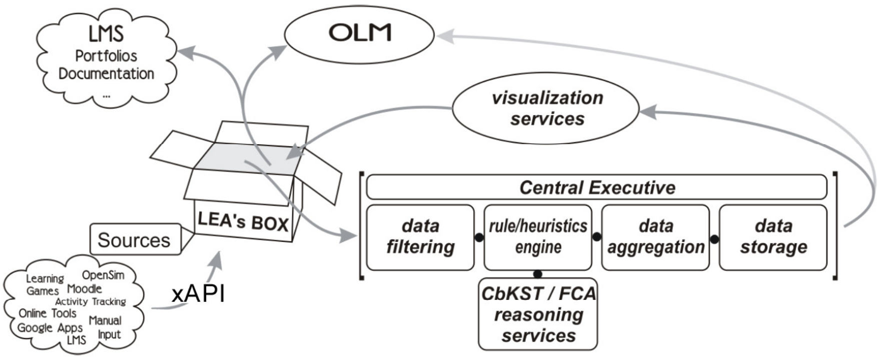

Fig. 1. Extended architecture of the Lea's Box system.

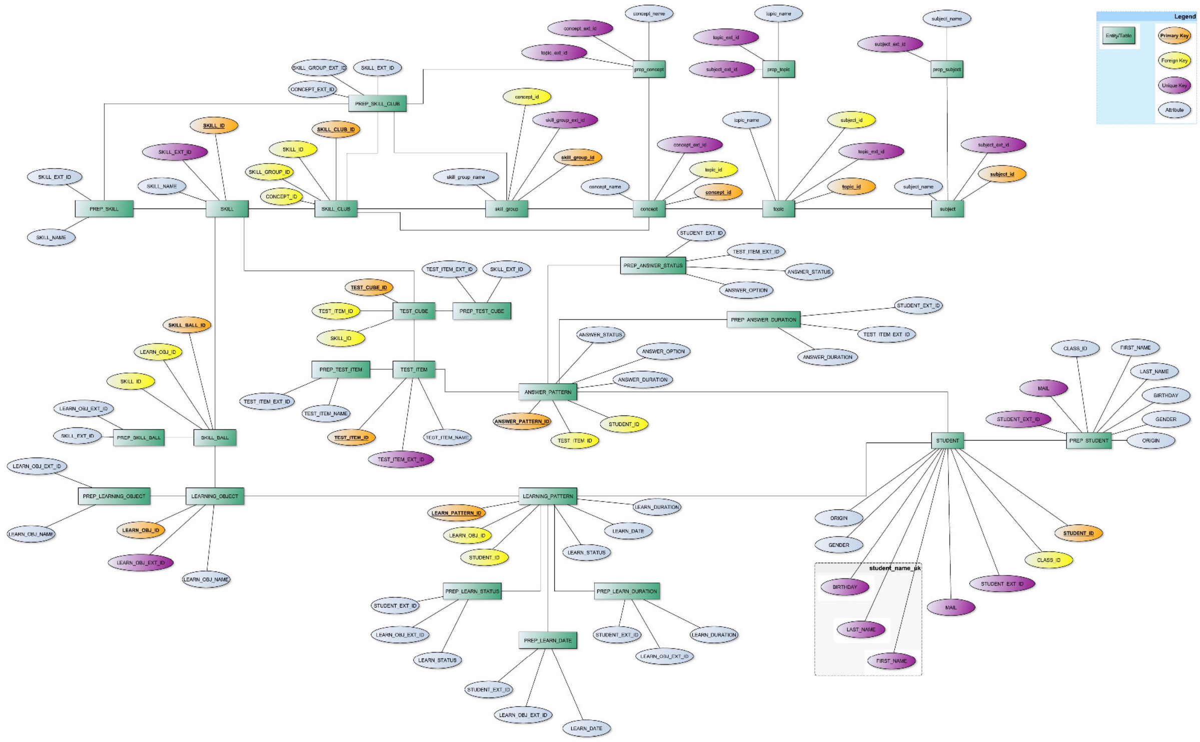

Fig. 2. Lea's Box data model.

Since other successful implementations of a data warehouse exist with MySQL, as shown by business practitioners (Anguanti, 2011; Sennhauser, 2016) educational institutions (Ingham, 2000), or have been integrated into big data solutions [18], the value of such solutions has been demonstrated. In addition, for a small project as Lea's Box, quick development, deployment and plug-in capacity are of utmost importance.

Within the database, a hybrid architecture with third normal form and a dimensional presentational area as suggested by Thusoo et al. (2010) are currently developed. The third normal form has been realized already. The hybrid approach was chosen to better fit the presently existing databases in the project. To make the database capable of high-speed data communication, a dimensional area with as few joins necessary as possible for queries, is in planning for the final phase of the project, to start with this summer. This will have to be weighed against the question of easier data integration from out-of-the-project sources like Moodle or inside-the-project sources like the Open Learner Model of the University of Birmingham (Bull et al., 2013) a partner in the project.

The final issue that was kept clearly in mind regarding the concept of the database was the potential barrier that a sophisticated extract, transform, loading (ETL) system could have posed for integration of data and for the budget of the project, both in time and financial resources. For this reason, a staging area (preparatory, or PREP, area) within the database was realized, that enabled it to write data into the database that possesses many NULL values and transform it within the database, without the need to lay another level of software between an outside source, the API and the services of the database.

2.2. Design

The overall design of the database is visible in Fig. 2. A snowflake structure design was deemed most appropriate. Differing from a design by the textbook, the database was built to have a four-fold center, which coincides with the possible angles of analyses. These are:

1. skills/competencies
2. test items/answers
3. content/learning accomplishments
4. personal data

Skills/competencies are central to the Lea's Box psycho-pedagogical structure that is competence focused, with a pronounced depth of skill hierarchies that can be more than 10 levels deep and might also be configured as tridimensional maps of complex relationships. The model should fit this layout. Personal data so far has been captured mainly to identify groups and control variables for analyses, but should be thought of as a slowly changing dimension that might lead to new insights once longitudinal data is put in. The difference between test items and learning accomplishments is that the former are defined by answer options that yield a (more) correct/incorrect outcome at some point, whereas the latter are characterized by a structure that differentiates between participation statuses alone.

The presentational area so far is work in progress. Although impressive access to information is possible even now, as can be seen from the section about temporal analysis in this paper, the presentation to an audience that is less technological savvy is still open. Several approaches are pursued, either by integration into established systems like the OLM or Moodle (in future, via the xAPI[4]) or by new solutions from the project, such as landscape visualization. For this part of the database that should have the capacity to act as a standalone service, processing data transferred via the xAPI, a dimensional model will be developed during the final months of the project in late 2016. This will make the queries faster and present a cube like data structure, being geared to the standards of Business Intelligence applications. This area will make the slicing and dicing of educational data available to administrators and practitioners alike.

The xAPI was selected as the connection most appropriate for the database, as it has demonstrated tremendous capacity for interoperability (Kitto et al., 2015), providing an easy, quick and secure transfer of educational data (Judrups, 2015).

2.3. Obtaining data

Between the preparedness of the structure and the results in a live school environment, there are huge obstacles that manifested in the quest of obtaining data. First of all, schools in Europe are, generally speaking, technology sparse organizational systems. This leaves much to be desired in the realm of data capture. And apart from what influential glossy technological magazines might fantasize about (Anderson, 2008), the age of data storage in the petabyte dimension still needs models, especially where hypotheses are central to not only making a process visible but making it understandable and in this way changeable. Barbera Gros, and Kirschner (2015) were dead-on when they stated that: "This perception of more time spent and higher workload [as a consequence of technology implementation]

[4] https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/

has been found to be related to a lack of knowledge of and abilities with respect to ICT for teachers and administrators." It will not do to hope for the rise of a "digital native", tech-savvy cohort, finding their place in educational organizations.

The question remains open of how to get schools to work together with researchers when the resources perceived to be scarcest by the practitioners are attention and time. Nor do school children leave a digital footprint that is easily accessible to scientists, nor should it be. Lessons learned from the project were that a tight-knit interpersonal relationship is necessary, which takes time to build up and that cooperation base firstly on persons and not organizations. One result of this is that, especially schools that are advancing in technological implementations, are feeling beleaguered by requests to dedicate more time of their professionals and students to undertakings that have no calculable return on investment or clear sustained future, being based on project-to-project funding.

Additional obstacles faced by the authors were strict and sensible rules and guidelines on data privacy and tracking. Higher education systems like universities or academies for professional have the ungrudgingly existent advantage of their students being of age or even customers that are in a contractual relationship. This results in a documentation that is both rich and deep in nature, if the institutions have the necessary funding. It is an answer to be found how these high standards of data privacy, clear-cut reduction to only the indispensable intrusion into the stream of information produced by students or teachers and advancements of epistemological meaningful results can be married.

To test the usefulness of structures and also to overcome the question of what is in for the schools targeted for study, data simulation might be of great use. Typically, after gaining confidence of a single teacher, implementation and wider spread of usage is mostly word-of-mouth propaganda within the organization, which, without belittling the importance of that to gain insights into school procedures in the field, is of limited use in the short time frame of most scientific projects. But with insights presented to the administration that are generated from the projected path of a few hundred students over the next few years, and are based on scientifically sound assumptions and axiomatic verifiable models, the path of entrance into educational data might be smoothed. However, so far there seems to be a lack of models to soundly compute individual student's data, especially in the area presented in the next section.

3. Temporal Measures as Distinct Features

As Barbera, Gros, and Kirschner's (2015) systematic literature analysis shows, time is both central and peripheral to educational research.

Its central status derives from the fact that ICT modifications of existing learning practices, usually desired to be an improvement or at least an amendment, are time-focused: They are constrained by the fact that they are "time-consuming" for students and teachers alike while they are heavily relying on the timely delivery of their content. Their advantages are also quite often of a temporal nature, as they promise an asynchronous availability of feedback, support or content as a benefit; additionally, a more speedy delivery through channels formerly not used for education promises a more focused and uninterrupted learning process, as in the case of flipped-classrooms via learning management systems or mobile applications.

Peripheral is time in educational research with the meaning of that it is seldom assessed as an end in itself. Most of the time, temporal measurements are used as control variables for other measures, e.g. retained knowledge, delivered content or degree of competency acquired. The educational community wants for research designs that focus on the usage of time, as proposed by a longitudinal research design which is event-centered, that [9] advocates.

Arising from the nature of the data processed so far, a longitudinal examination was not possible. However, the database can considerably facilitate a combined variable-event-approach by providing both information on temporal factors and of the choices, events and order of tasks or activities that were undertaken. This brings the goal to exhaustively use the temporal data of students closer into reach of researchers, practitioners and organizational specialists.

Apart from the availability of an otherwise neglected resource, data about time is crucial for pedagogical success in other ways, too. As an example for the deployment of temporal resources, of which there are way too less studies for a thorough model, Langa (2013) found that teachers in tertiary education generally lacked instructions for their students about how much time they should allocate to self-study time, with sometimes quite negative results. Although there is not a clear relationship between time dedicated to self-study and educational success, spending more time on individual assignments lead to higher achievements in this study. This finding is corroborated by the extensive study of Morris, Finnegan, and Wu (2005), which tracked students for 1.5 years in their online coursework. A clear distinction arose between the three groups of (successful) completers and the non-completers of the courses, most heavily correlating with the amount of time spent on discussion pages and the amount of content/discussion pages read.

These findings point to the close connection between engagement, resource-allocation and achievement. However, as remarked above, without a more fine-grained analysis that takes into account order of activity and also more acute measures, the predictive value and the feedback to the author/teacher of a course remains foggy at best. One should not assume that "data speaks for itself", but remain aware of the close limitations that raw and coarse-grained data entails (Long and Siemens, 2011).

When considering middle or high school children who are a fair bit more dependent on exact and stepwise instructions, one cannot stress enough that temporal measures of tasks handed out for self-study are important, as results of homework are generally graded or given feedback upon after a relatively long time has passed, sometimes weeks.

3.1. Data analysis

The data that was analyzed for the purpose of this paper, as a proof of concept, stemmed from an extra-curricular activity in middle school biology, focusing on evolution. Thus, the activity is under more intense scrutiny for its benefit by both teachers and students as it lacks the authority of the official textbook for the course. The provider of the course, however, only gives feedback on "traditional" measures: content viewed/(non-)completed, (un-)successful answer to test items, aggregable by group only. To complement the great content with further use for the learners and teachers, the logged data was examined completely with regard to creating additional value for further implementations.

The temporal analysis was deemed most useful, since it a) is directly available after the completion of a segment of the course, is b) easily comparable across students and groups and teachers and c) provides information that is not obtainable for classroom or homework activities unless huge efforts are taken or costly technology is available.

The data set consists of the log-data of a total of 105 students that viewed three information sources (learning assignments) which were presented to them online. After the familiarization with the content, the students answered five test items which were logged as either "correct", "incorrect" or "not completed". The time of viewing per learning assignment and the time until an answer was given were also monitored for each individual student.

All student data was anonymized before it was sent to the database by the vendor of the online course. The only information disclosed were the grade of the students, a group membership on class level, as given by the teacher, and their location on state level. All students were in the 7th grade and were from Texas, USA. The whole set comprised 6 groups of Science classes. Even though the number of students per group was unevenly distributed, with rather low numbers per group, a group-wise analysis was conducted for purpose of demonstration.

Keeping in mind that the analysis should be something that is easily understandable without statistical background and should also be simple to visualize, (un-)completed assignments and a temporal analysis of completed assignments was conducted firstly. The limitation on completed assignments arose from the circumstance that, sadly, the vendor did not log temporal information on uncompleted assignments.

It can be discerned from Table I that even though the average time of completion of the learning assignments per group is evenly distributed, there is a considerable difference in the deviation from this mean. This would be valuable information for a teacher, pointing to difficulties and to different learning strategies, as well as time spent on homework assignment, an insight normally not available.

TABLE I. TEMPORAL ANALYSIS OF THE ASSIGNMENTS

Group	N[a] of students	Students' achievements		
		Sum of CA[b]	*Sum of UCA[c]*	*Average time of completed items (SD)[d] [s]*
Science 7, 2nd	15	45	0	519.31 (353.46)
Science 7, 4th	15	44	1	459.52 (285.28)
Science 7, 6th	23	62	7	497.69 (417.75)
Science 7, 7th	15	43	2	507.09 (469.79)
Science 7, 8th	16	48	0	513.90 (421.91)
Science 7, 9th	21	53	10	542.04 (579.89)

a. Number
b. Completed Assignments
c. Un-completed Assignments
d. Standard Deviation

In contrast to the learning assignments, for the test items presented in Table II, there exists temporal log-data for both correct as well as incorrect answers. Again, easily discernible are pronounced differences between the temporal patterns of the items. Items 2, 3 and 4 have low deviance regarding correct answers, which speaks for an established knowledge for these tasks. Contrasting that is the noticeable difference for test items 1 and 5. Although item 1 was the easiest, it had the biggest deviance for correct answers, on a level with item 5, which was the hardest. This contrasts with the intuitive assumption "easiest is quickest (for everyone)".

TABLE II. TEMPORAL ANALYSIS OF THE ASSIGNMENTS

Test item	Sum of correct (incorrect) answers[a]	Students' achievements	
		Average time of correct items (SD)[b] [s]	*Average time of incorrect items (SD)[b] [s]*
1	98 (1)	19.85 (75.85)	137.00 (0.00)
2	96 (3)	17.60 (30.67)	11.33 (5.73)
3	94 (5)	13.80 (20.87)	7.00 (2.28)
4	91 (8)	18.36 (27.89)	43.00 (51.51)
5	88 (11)	60.55 (62.43)	50.00 (59.47)

a. Missing amount to 105 where non-reported values of six students, either due to technical issues or non-completion of the test item
b. Standard Deviation

Interestingly, a different pattern can be seen for the incorrect answers. Test items 2 and 3 are both quickly and evenly distributed answered incorrectly. However, both items 4 and 5 take considerably longer when they are answered incorrectly. This markedly contrasts for item 4, where incorrect answers took more than twice the time of correct answers.

4. Further Insights

The initial insights gained from the data underline the need for further inquiry into the temporal characteristics of pedagogical tasks. With scarcity of time constantly named as one of the top issues by teachers, students and parents, this irrecoverable, once spent, resource cannot be neglected when thinking about the advantages one could gain from an information-(en)rich(ed) classroom or learning system. Additional information about the time spent per task or content also gives rich feedback and substantiates discussions about learning patterns, styles and problems that arise from a forced (a) synchronicity of time spending in both formal and informal settings. For this, a research and monitoring approach that takes into consideration the importance placed on order, sequence and paths by Reimann (2009) can reap low-hanging fruits of data for increased relevance of evidence-based pedagogical efforts. To garner these rich data fields the overcoming of barriers, both in psychological/socio-cultural and technological /logistical areas seem a central obstacle. Time is intimate as it is unequivocally up to individual choices how it is spent. When and for which length tasks are undertaken speaks as much, sometimes even more, about consensus or dissent regarding educational matters than care, intensity and efforts shown in results. Therefore, as with all data, especially from minors, a careful balance has to be sought between increased empirical substance of decisions and the privacy and space for necessary mistakes or more vital tasks at hand for students.

5. Conclusions

Learning Analytics solutions that focus on less well-defined settings such as analyzing data of a specific e-learning system, where (a) all elements are designed tailored to this system and where (b) all activities occur in a digital environment, can easily work up the data and provide analyses to the educators. Learning Analytics solutions that aim at accommodating the various needs of different educational settings ranging from primary school to higher education up to workplace learning are confronted with much greater challenges. There are various, in parts incompatible, competence models, there are extremely diverse demands on a dynamic grouping of learners, there are diverse types of data life cycles, and there is an extremely broad range of data formats that are spilled into the system. Lea's Box is a project that is working in such complex scenario. The project's vision is to support teachers in their daily educational routines and provide them with means of using data and data analytics to optimize their teaching. In the typical school context, the data we can fall back on are thin, unfortunately. Usually, in school settings learning data are primarily generated by teachers' manual assessments and marking of students' works. Additionally, students use diverse learning apps, perhaps they accomplish an exercise in the computer lab every now and then, at best, homework is made in some electronic form. Such data basis is far from being coherent or complete. In far most cases they are not brought together as basis for educational decisions of a single teacher, even more seldom are scenarios where data 'owned' by multiple teachers are brought together. The reasons are quite clear: Most tools used in education do not provide a suitable data export and most Learning Analytics solutions cannot handle diverse and incomplete data. Lea's Box attempts to provide such solution. On the basis of open APIs and import features, we want to equip educators with a system that takes every bit of information that it can get and provides the best possible analytics for the teacher.

This, of course, is not trivial. In the past we saw a large number of attempts to develop educational standards and to achieve interoperability. Almost all failed in achieving a substantial uptake in education. Today, a good portion of tools used in school settings are not even primarily educational; teachers are using Google docs for homework, Dropbox for file sharing, or Minecraft for practicing. The way Lea's Box addresses the problem is to start from competence frameworks. We see competencies as the main entities in any educational setting; it's always about improving the knowledge, the skills, the aptitude of learners. We define competencies, we try to identify the relationships among them and we try to find a natural course of learning or development (Kickmeier-Rust, Bull, and Albert, 2016). Having such beacons, we provide an open system that allows linking all sorts of activities, all sorts of data as evidences for the competencies. By this means we can maximize the data sources that are contributing to learning analytics and support teachers in a competence-focused way in by providing open and transparent models of learners.

The essential foundation for such system is a data model that s capable of dealing with the described diversity and that is capable of analyzing the data in a high-speed and scalable way. In this paper we sketched the approach developed in the Lea's Box project. In ongoing user studies we are bringing the solution into practice and evaluate the effectiveness and adequacy of the approach. The studies run in various countries (i.e., Austria, Germany, the Czech Republic, Turkey, and the UK) and at various school levels. Our first experiences show that the described 'data analyses warehouse' meets these expectations and provides a strong and high-speed approach to 'ill-defined' learning analytics settings.

Acknowledgements

The Lea's Box project that is responsible for the work presented in this paper is supported by the European Commission under the 7th Framework Programme, contract number 619762. This document does not represent the opinion of the EC and the EC is not responsible for any use that might be made of its content.

References

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, June 23, 2008, <http://www.wired.com/science/discoveries/magazine/16-07/pb_theory>.

Anuganti, V. (2016). Designing Scalable Data Warehouse with MySQL, presentation at the *Oreilly MySQL Conference*, 2011. (http://www.slideshare.net/vanuganti/designing-scalable-data-warehouse-using-mysql, accessed at 2016-05-27).

Barbera, E., Gros, B., & Kirschner, P. A. (2015). Paradox of time in research on educational technology. *Time and Society*, vol. 24(1), pp. 96-108

Bull, S., Ginon, B., Boscolo, C., and Johnson, M. D, (2016). Introduction of Learning Visualisations and Metacognitive Support in a Persuadable Open Learner Model. In *Proceedings of the workshop Learning Analytics for Learners at LAK'16 conference*, 26-29 April, 2016, Edinburgh, UK, 2016.

Ingham, J. (2000). Data Warehousing: A Tool for the Outcomes Assessment Process. *IEEE T. Educ*., vol. 43 (2), pp. 132-136.

Idreos, S., Groffen, F., Nes, N., Manegold, S., Mullender, S. & Kersten, M. (2012). MonetDB: Two Decades of Research in Column-oriented Database Architectures. *Quart. Bull. Comp. Soc. IEEE TC Data Eng.*, vol. 35, pp. 40-45.

Jo, I.-H., Kim, D. & Yoon, M. (2014). Analyzing the Log Patterns of Adult Learners in LMS Using Learning Analytics. In *Proceedings of LAK '14*, March 24 - 28 2014, Indianapolis, IN, USA, pp. 183-187.

Judrups, J. Zandbergs, U., Arhipova, I., & Vaisnore, J. (2015). Architecture of a Competence-Based Human Resource Development Solution, *Procedia Comput. Sci.,* vol. 77, pp. 184-190.

Langa, C. (2013). Management of time resources for learning through individual study in higher education. *Procedia – Soc. Behav. Sci.*, vol. 76, pp. 13-18.

Long, P. & Siemens, G. (2011). Penetrating the Fog – Analytics in Learning and Education. *Educause Rev*., vol. 46(5), pp. 31-40, September/October 2011.

Kickmeier-Rust, M. D., Bull, S., & Albert, D. (2016). LEA's BOX: Practical Competence-oriented Learning Analytics and Open Learner Modeling," in *Proceedings of the workshop Learning Analytics for Learners at LAK'16 conference*, 26-29 April, 2016, Edinburgh, UK.

Kitto, K., Cross, S., Waters, Z., & Lupton, M. (2015). Learning Analytics beyond the LMS: the Connected Learning Analytics Toolkit. *Proceedings of LAK'15*, Pughkeepsie, NY.

Masud, A. & Huang, X. (2012). An E-learning System Architecture based on Cloud Computing. *World Acad. Sci. Eng. Technol.*, vol. 62, pp. 74–78.

Morris, L. V., Finnegan, C., & Wu, S. (2005). Tracking student behavior, persistence, and achievement in online courses. *Internet and Higher Education*, vol. 8, pp. 221-231.

Rayon, A., Guenaga, M., & Nuñez, A. (2014). Ensuring the integrity and interoperability of educational usage and social data through Caliper framework to support competency-assessment. In *Proceedings of the 2014 IEEE Frontiers in Education Conference,* pp. 2775-2783.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *Int. J. Comp.-Supp. Coll.,* vol. 4, pp. 239-257.

Sennhauser, O. (2012). *Data Warehouse (DWH) with MySQL, original title, "Data Warehouse (DWH) mit MySQL.* Presentation at the DOAG Konferenz 2012. (http://www.slideshare.net/shinguz/data-warehouse-dwh-with-mysql, accessed at 2016-05-27).

Sonwalkar, N. (2013). *The First Adaptive MOOC: A Case Study on Pedagogy Framework and Scalable Cloud Architecture—Part I.* MOOCs Forum, vol. 1(P), 22-29.

Thusoo, A. Shao, Z., Anthony, S., Borthakur, D., Jain, N., Sarma, J. S., Murthy, R., & Liu, H. (2010). Data warehousing and analytics infrastructure at Facebook, Proceedings of the 2010 *ACM SIGMOD Int Conf Manage Data*, pp. 1013-1020, 2010.