(2016). Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics*, *3*(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

Editorial: Datasets for Learning Analytics

Stefan Dietze
L3S Research Center
University of Hannover, Germany
dietze@l3s.de

George Siemens
LINK Research Lab
University of Texas Arlington
gsiemens@gmail.com

Davide Taibi
Institute for Educational Technology
National Research Council of Italy
davide.taibi@gmail.com

Hendrik Drachsler
Welten Institute
Open Universiteit, The Netherlands
Hendrik.Drachsler@ou.nl

ABSTRACT: The European LinkedUp and LACE (Learning Analytics Community Exchange) project have been responsible for setting up a series of data challenges at the LAK conferences 2013 and 2014 around the LAK dataset. The LAK datasets consists of a rich collection of full text publications in the domain of Learning Analytics and Educational Data Mining. The LAK dataset offers publicly available, machine-readable versions of research articles from the Learning Analytics and Educational Data Mining communities in various formats, where the main goal is to facilitate research, analysis, and smart explorative applications. Based on the insights gained from these data challenges, the idea was born to make more Learning Analytics data sets publicly available for researchers to get to a more open access data-driven research community within Learning Analytics. With this special section, we publish four data sets that answered the call for data sets by the journal. It is our vision to collect more data sets like these in this initial collection and reward their creators through citing the datasets and connecting new research outcomes to them.

Keywords: Datasets, open access, science 2.0, LACE project

1 INTRODUCTION

Although learning analytics (LA) is applied increasingly in various learning and education settings, it still lacks publicly available and interoperable research datasets. Though LA research is rapidly advancing,

(2016). Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics*, *3*(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

the community lacks a sufficient foundation of open, reusable, publicly available datasets that would allow the reproduction, experimental evaluation, and benchmarking of algorithms, methods, and tools in the learning analytics area. Given the data-centric nature of the analytics in education, availability of sound benchmarks and large-scale datasets is a key enabler for maturing the field, assessing the state of the art, and comparing competing approaches and methods.

The European LinkedUp and LACE (Learning Analytics Community Exchange) project have been responsible for setting up a series of data challenges (Drachsler et al., 2014b) at the LAK conferences 2013 and 2014 (d'Aquin, Dietze, Drachsler, Herder, & Taibi, 2013; Drachsler et al., 2014a) around the LAK dataset. The LAK Dataset[1] means to provide an unprecedented and publicly available resource for structured data about learning analytics research — i.e., the actual research works as captured in scholarly papers from the community. The respective learning analytics data used by such research works is hard to discover and reuse. Data is either not provided, for of a variety reasons, or it is spread across distributed and disparate endpoints.

This special issue collects such datasets, i.e., data that arises from actual learning processes in any domain that is used within LA research and practice. This might involve datasets that facilitate further methodological development or that are of direct utility for learning analytics, including datasets that:

- Enrich learning analytics or educational data mining scenarios
- Help evaluation of learning analytics, educational data mining, or related tools and methods
- Provide a challenging test case for algorithmic or model development
- Provide a scenario for visualization techniques
- Express data in an existing or emerging data standard, either as a collection of reference examples or as a test case for interoperability (i.e., to test whether sufficient meaning can be reconstructed without knowledge of the source system)
- Specific combinations of datasets designed to test a particular theory or hypothesis, accompanied by explanatory rationale and any research findings already derived, to aid replication studies and theoretical development

To facilitate reuse of datasets, all presented datasets in this special issue comply with the following criteria:

- All data needs to be made available under open license terms (e.g., CC-BY, Open Data License) available for reuse by third parties
- The dataset provider needs to hold all rights to share the data publicly on the Web
- Data needs to be accessible online and preferable as a dump or via a public HTTP-accessible API

[1] http://lak.linkededucation.org

(2016). Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics*, *3*(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

or SPARQL endpoint

- Data needs to be accessible in standardized serializations and formats, such as XML, CSV, JSON, or RDF. Data should be complemented with a description of the fields, a schema file, and/or vocabulary description

In the current special section, we have collected four datasets from various backgrounds:

1. Stamper, J., & Pardos, Z. A. (2016). The 2010 KDD Cup competition dataset: Engaging the machine learning community in predictive learning analytics. *Journal of Learning Analytics, 3*(2), 295–299.

Description: In the spring of 2010, the Association for Computing Machinery (ACM) Special Interest Group on Knowledge Discovery and Data-mining (KDD) selected an educational technology dataset for its annual competition. The competition, titled "Educational Data Mining Challenge," tasked participants with predicting the correctness of student answers to questions within an Intelligent Tutoring System (ITS) from The Cognitive Tutors suite. PSLC DataShop hosted this challenge and included data provided by Carnegie Learning Inc., producers of The Cognitive Tutors. Consisting of over 9GB of student data, this was the largest KDD Cup dataset up to that time. The competition brought in 655 competitors submitting 3,400 solutions. Five years later, the competition dataset has been the most often cited from an educational technology platform.

2. Papoušek, J., Pelánek, R., & Stanislav, V. (2016). Adaptive geography practice data set. Journal of Learning Analytics, *3*(2), 300–304.

Description: A dataset on student learning of geography facts in an open online system — slepemapy.cz. The data set has a simple format with intuitive interpretation. At the same time, it offers rich possibilities for modelling and analysis — for example, prior knowledge, forgetting, or response times. The data set is based on an open education system — an open source project freely available online — with available description of algorithms used. Researchers can thus try the system themselves before using the data set and inspect the details of its realization. This is in contrast with many current education data sets whose origin is not completely clear or easily inspectable.

3. Vozniuk, A., Holzer, A., & Gillet, D. (2016). Peer assessment dataset. *Journal of Learning Analytics, 3*(2), 305–307.

Description: Peer assessment is seen as a powerful supporting tool to achieve scalability in the evaluation of complex assignments in large courses, possibly virtual ones, as in the context of massive open online courses (MOOCs). However, the adoption of peer assessment is slow, due in part to the lack of ready-to-use systems. Furthermore, the validity of peer assessment is still under discussion. In order to tackle some of these issues, a dataset is presented containing the assessment of student submissions by student peers and by instructors during our Social Media course with 60 Master's level university students. The dataset allows for training and testing algorithms that predict the grades of instructors based on the grades of student peers.

(2016). Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics*, *3*(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

4. Tabuenca, B., & Börner, D. (2016). Noise in classrooms data set. *Journal of Learning Analytics, 3*(2), 308–312.

Description: A dataset comprised of noise samples collected with a mobile device during 26 sessions of a technology class at a secondary level. The dataset includes rich metadata that can facilitate correlation with further studies, namely type of session (i.e., traditional face-to-face lecture, collaborative workshop session, individual computer session), number of students participating in the session, percentage of male/female students, mean age of the students, timestamp when the sample was collected, language of the session, and country, city, and location where it took place. The data is shared in different formats to facilitate its management across platforms.

For the future of learning sciences research, we envision a learning analytics data repository to host various data sets from all kinds of educational domains, systems, and contexts (e.g., kindergarten, K–12, higher education, and workplace learning, but also MOOCs, LMSs, digital games for learning, online communities, and physiological sensor data like eye-tracking or motion capture traces). Datasets of relevance could include data about cognitive development, social learning, discourse progression, network interactions, learning paths through courses, competency completion, help-seeking behaviour, and distributed multi-spaced interactions. While such primary data about learning processes are of central importance, complementary data gathered through surveys, for instance, about learner demographics, background knowledge, goals, perceptions, experiences, and attitudes would be of relevance.

Such a central learning analytics data repository could contribute to maturing the evidence for learning analytics and gain comparable results to natural science communities (Kalz & Drachsler, 2016. This would empower future research, starting with existing data sets and published learning analytics approaches, and building upon the insights gained from the previous studies to contribute new insights to a body of knowledge on learning analytics. With such an approach, we would also acknowledge the creators of representative data sets, expressing those acknowledgements through citation by other researchers in the community.

Having said this, we want to keep the line of submission of educational datasets open at the *Journal of Learning Analytics* in order to continuously collect additional datasets and enable learning analytics researchers to be known not only for their research, but also for the datasets they have created.

ACKNOWLEDGMENTS

The European Commission Seventh Framework Programme funds the Learning Analytics Exchange (LACE) project: Grant Number 619424

REFERENCES

d'Aquin, M., Dietze, S., Drachsler, H., Herder, E., & Taibi, D. (2013). *Proceedings of the CEUR Workshop*

(2016). Editorial: Datasets for Learning Analytics. *Journal of Learning Analytics*, *3*(3), 307–311. http://dx.doi.org/10.18608/jla.2016.32.15

(LAK Data Challenge 2013) held at the Third Conference on Learning Analytics and Knowledge (Vol. 974). Leuven, Belgium: CEUR-WS.org Retrieved from http://ceur-ws.org/Vol-974/

Drachsler, H., Dietze, S., d'Aquin, M., Herder, E., & Taibi, D. (2014a). *Proceedings of the CEUR Workshop (LAK Data Challenge 2014), held at the 4th International Conference on Learning Analytics and Knowledge* (Vol. 1137). Retrieved from http://ceur-ws.org/Vol-1137/#workshop1

Drachsler, H., Stoyanov, S., d'Aquin, M., Herder, E., Guy, M., & Dietze, S. (2014b). An evaluation framework for data competitions in TEL. In C. Rensing, S. de Freitas, T. Ley, & P. J. Munoz-Merino (Eds.), *Proceedings of the 8th European Conference on Technology Enhanced Learning* (EC-TEL '14), (LNCS Vol. 8719, pp. 70–83). http://dx.doi.org/10.1007/978-3-319-11200-8_6

Kalz, M., & Drachsler, H. (2016). The MOOC and learning analytics innovation cycle (MOLAC): A reflective summary of ongoing research and its challenges. *Journal of Computer Assisted Learning, 32*(3), 281–290. http://dx.doi.org/ 10.1111/jcal.12135