Forging new pathways of research and innovation in open and distance learning:
Reaching from the roots
Proceedings of the 9th European Distance and E-Learning Network Research Workshop, 2016
Oldenburg, 4-6 October, 2016
ISBN 978-615-5511-12-7
ISSN: 2707-2819
doi: https://doi.org/10.38069/edenconf-2016-rw-0026

SCALING LEARNING ANALYTICS: THE PRACTICAL APPLICATION OF SYNTHETIC DATA

Alan M. Berg, Stefan T. Mol, Gábor Kismihók, University of Amsterdam, Netherlands,
Niall Sclater, Sclater Digital Ltd, United Kingdom

Summary

This case study is based on experiences gained during the running of a two-day data hackathon around large scale Learning Analytics infrastructure at the LAK16 conference. The main conclusion is that there will be a significant demand for realistic synthetic data to support the development of large scale infrastructures.

Synthetic data overcomes ethical barriers to sharing large data sets between different (parts of) organizations. Properly simulated synthetic data can be leveraged to fine tune algorithms deployed within the field of Learning Analytics. This data driven approach lowers the risk of accidental disclosure and bypasses limitations rightfully imposed due to legal and/or ethical constraints associated with real student data. The application of synthetic data to performance testing allows universities to develop highly scalable infrastructure in parallel to developing central data governance practices.

This short paper explores the conformance testing of Learning Record Stores (LRS – secure locations to store and query student digital traces), discusses the implications for Universities around a specific set of xAPI recipes (Berg, Scheffel, Drachsler, Ternier, & Specht, 2016) and generalizes practices for the acceleration of large scale deployments of LA infrastructure. The authors argue that by applying a standardized set of synthetic data based on a peer reviewed synthetic data generator, universities will find it easier to develop reliable recipes for digital learner traces. Consistent data storage across university boundaries will subsequently enable the benchmarking of algorithms that consume student digital traces and support the generation of predictive validity evidence across university boundaries. Thus universities can compare the value of their algorithms relative to other universities and consistently apply algorithms when students transfer.

The relevance of synthetic data in Learning Analytics

Synthetic data, also known as simulated data, has been heavily researched and successfully applied across a broad range of scientific fields. Berg, Mol, Kismihók, & Sclater (2016) have previously discussed the application of synthetic data within the field of Learning Analytics.

Based on experiences developed through the UvAInform project (Kismihók & Mol, 2014) at the University of Amsterdam, that involved the development and conduct of seven pilot experiments built onto a central Learning Record Store, the following reasons to deploy synthetic data were identified:

Synthetic data can:

- Enable the parallel development of data governance processes while developing infrastructure without exposing real data.
- Populate performance tests with realistic load patterns.
- Potentially support the development of benchmarking across organizations without the release of a real data set
- Populate a training environment without ethical or security risks.
- Enforce best practices by codifying conformance tests.

LA infrastructure

The LAK 16 data hackathon (2016) brought together two organisations leading the way worldwide in developing open architectures for learning analytics. Jisc (2016) represents the UK further and higher education sector and is a not-for-profit organisation for digital services and solutions, and Apereo (2016) is a foundation that supports the development of Higher Education open source software. The hackathon put the growing ecosystem of learning analytics products such as learning records stores, learning analytics processors, dashboards, consent systems and student apps through their paces with synthetic big data from learning management systems, student record systems and other sources.

The motivation was to test the interoperability of the various tools, and to integrate new data sources, predictive models, and third party products to help to reassure institutions that they are not going to be locked into proprietary learning analytics systems and that they will be able to select the best products for their needs. xAPI, an emerging standard for the exchange of learning records, was the basis of efforts to develop new profiles, which are also known as *recipes* (Scheffel, Ternier, & Drachsler, 2016) and applications for learning analytics.

Jisc provided the infrastructure based on their national learning analytics architecture (Sclater, Berg & Webb, 2015) and test data, which was then converted to xAPI calls via an open source tool running in specially crafted test plans. The data that drove the test plans was synthetic, however, the distributions were based on captured digital traces. The main technical artifact was a secure repository for digital traces mostly from students. The repository receives the traces via the xAPI protocol and is queried by the same protocol.

Figure 1 displays one of many possible realistic infrastructures. Digital traces are captured through javascript libraries in web pages. Student Dashboards query the LRS and so does a data warehouse. Within the data warehouse, Machine Learning Algorithms are applied that

ISBN 978-615-5511-12-7

generate predictions about the students. Dashboards can then query an Educational API about the results from the algorithms. The Educational API queries the data warehouse.

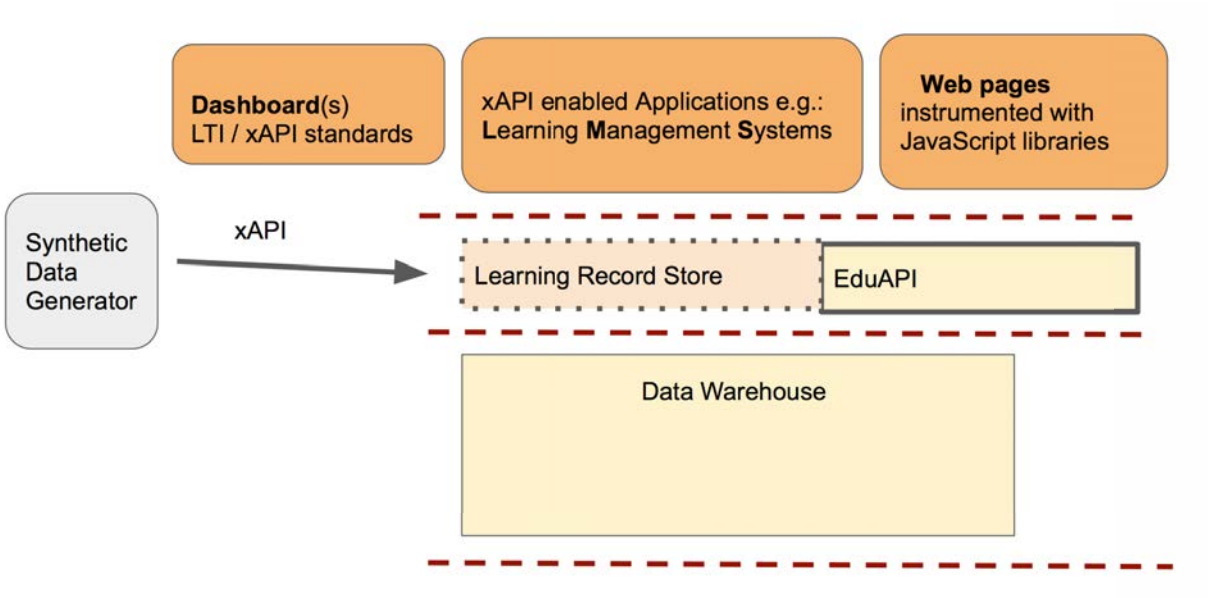

Figure 1. Simplified LA Infrastructure

By tasking a Synthetic Data Generator (SDG) to populate the Learning Record Store one can simulate the student activity and test the conformance and performance of the infrastructure without real users and their "real" xAPI traces. When necessary the SDG can also generate synthetic Student Information System input for the data warehouse. The data that drove the SDG for the hackathon was fully artificial based on a dataset released by Jisc which resides in the Hackathon github repository (2016). The Jisc dataset was itself synthetic based on real usage. However, as more datasets are made available, the data can be analysed, real distributions of usage deducted and then encapsulated directly into the test plans. In other words, the more datasets, the better the fidelity and generalizability of the data generated by the SDG.

The authors choose Jmeter (2016) as the basis for the Synthetic Data Generator. The application is open source software, a 100% pure Java application designed to load test functional behaviour and measure performance. Jmeter was chosen because it:

- is open source, hence the code is open to review, re-use, and alteration;
- has a rich feature set so one can, out of the box, write test plans that deliver complex datasets;
- is extendable through a scripting language; so if necessary one can extend the richness of datasets programmatically;
- is a mature product with a considerable amount of documentation and usage;
- has a drag and drop GUI which simplifies test plan creation;
- is data driven, e.g. one can modify the tests simply by changing CSV files;

ISBN 978-615-5511-12-7

- has a master slave architecture so can scale to extremely high loads e.g. one Jmeter application can run many other Jmeter applications across a server park;
- enables the repurposing of conformance tests as performance tests.

Jmeter runs test plans which are designed in a GUI interface and saved in XML format. Figure 2: displays a hackathon test plan. On the left hand side of the GUI are elements that perform one specific task well. On the right hand side is the configuration for a highlighted element. The majority of the elements in the test plan, specifically the elements in the top left corner read in CSV files containing data about user names, course ids. The data is necessary to send meaningful xAPI requests with a diversity of detail to the Learning Record Store. The element highlighted converts the one line of inputted data into a well formed HTTP REST request. Jmeter can run many requests at the same time, when necessary generating a considerable amount of load even from one machine. As noted above, one of the motivations for choosing Jmeter was the wealth of documentation. A Google search with the term "how to do performance testing using Jmeter" returns around 300,000 results.

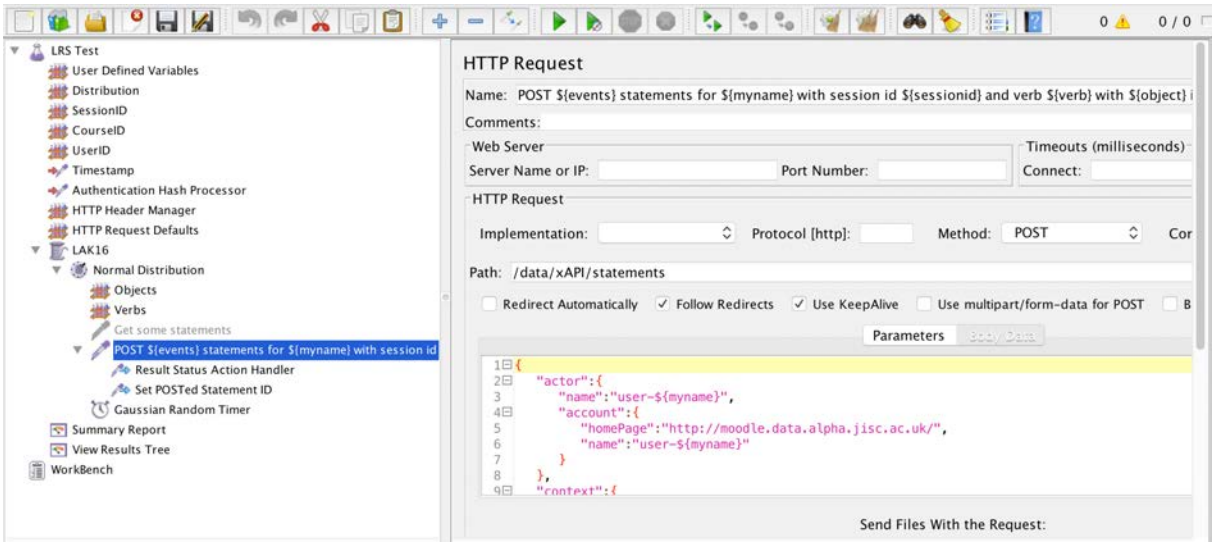

Figure 2. Jmeter test plan

In return for the mild complexity, the test plan generated a rich set of xAPI recipes that were subsequently used to populate a number of different commercial Learning Record Stores during the hackathon. The Learning Record Stores followed the xAPI specification (2016), but responded subtly differently to the requests. The conformance tests helped isolate a number of unexpected behaviours which were then rapidly repaired during the hackathon.

A single Jmeter instance also populated the main LRS with over 400,000 records within a 30-minute period, delivering confidence that the test infrastructure could withstand unexpected load. The quality of the data was limited by the lack of variety in the Jisc dataset. As usage increases on the Jisc experimental national infrastructure the authors expect the richness of xAPI recipes to also increase and thus motivate further improvement in the open-sourced test plans. The test plans did not include population of the data warehouse, however Jisc have released examples of the necessary format, see Jisc data (2016).

ISBN 978-615-5511-12-7

If the same test plan is applied across organizational boundaries then the assumption is that the same xAPI recipes will also be applied across boundaries, leading to greater standardization and consistency. For example, one could consider deliberately limiting the conformance tests to work only with the Dutch xAPI Specification for Learning Activities (DSLA) or a formalized international equivalent. If the SDG fails to generate the data needed for a given dashboard or algorithm within the target infrastructure, then it is clear that the infrastructure is not conformant and would require subsequent effort for other organizations to share data and perform joint research on that data.

Roadmap

The SDG is at an early stage of development. The code resides in a public Hackathon github repository (2016) open to review, modification, and extension. To improve the initial simplistic test plans requires the gathering of requirements, the exploration of real data, and an agreement on which xAPI recipes to apply.

The wish list of improvements includes:

- An improved generation of distributions of xAPI recipes per student cohort. For example, finer grained division over a wide range of student cohorts.
- A complete set of conformance tests for Learning Record Store vendors.
- A thorough application of xAPI recipes intended to be consistent across organizations.
- Documentation for the deployment of the SDG.
- Extension of the SDG to standardized SIS data formats.
- The development of a benchmark for LA algorithms based on SDG data.

Conclusions

From the experiences garnered from the UvAInform project and the LAK16 data hackathon the authors argue that by applying a standardized set of synthetic data based on a peer reviewed SDG, universities will find it easier to develop consistent recipes for digital learner traces. Consistent data storage across university boundaries will later enable the benchmarking of algorithms that consume student digital traces supporting the insurance of predictive value across University boundaries. Thus universities can compare the value of their Learning Analytics algorithms relative to other universities and apply consistently the algorithms as students transfer.

An example SDG was provided for the LAK16 hackathon. The tooling is well documented and open source. Although basic the test plans are free to use and update.

ISBN 978-615-5511-12-7

References

1. Apereo (2016). *About.* Retrieved from https://www.apereo.org/content/about
2. Berg, A. M., Mol, S. T., Kismihók, G., & Sclater, N. (2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics, 3*(1), 107–128. http://doi.org/10.18608/jla.2016.31.7
3. Berg, A., Scheffel, M., Drachsler, H., Ternier, S., & Specht, M. (2016). The dutch xAPI experience. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge – LAK '16, New York*, 544–545. New York, USA: ACM Press. http://doi.org/10.1145/2883851.2883968
4. Hackathon github repository (2016). *hack-at-lack16.* Retrieved from https://github.com/AlanMarkBerg/hack-at-lack16
5. Jisc (2016). *About us.* Retrieved from https://www.jisc.ac.uk/about
6. Jisc data (2016). *'University of Jisc' Sample Data Set.* Retrieved from https://github.com/jiscdev/learning-analytics/blob/master/sample_udd_data.md
7. Jmeter (2016). *Apache JMeter.* Retrieved from http://jmeter.apache.org
8. Kismihók, G., & Mol, S. T. (2014). *Barriers to adoption for learning analytics at a Dutch University.* Presented at the Learning Analytics Summer Institute, Utrecht, the Netherlands. Retrieved from https://lasiutrecht.files.wordpress.com/2014/06/uvainform-presentation-lasi-utrecht-2014.pdf
9. LAK 16 (2016). *Jisc / Apereo Learning Analytics Hackathon.* Retrieved from http://lak16.solaresearch.org/?page_id=204
10. Scheffel, M., Ternier, S., & Drachsler, H. (2016). *The Dutch xAPI Specification for Learning Activities (DSLA) – Registry.* Retrieved from http://bit.ly/DutchXAPIreg
11. Sclater, N., Berg, A., & Webb, M. (2015). Developing an open architecture for learning analytics. *Proceedings of the 21st Congress of European University Information Systems (EUNIS 15), Dundee,* 303–313. Dundee, Scotland: European University Information Systems Organisation. Retrieved from http://www.eunis.org/wp-content/themes/eunis/assets/EUNIS2015_Book_of_Abstracts.pdf
12. xAPI specification (2016). Produced by the Experience API Working Group in support of the Office of the Deputy Assistant Secretary of Defense (Readiness) Advanced Distributed Learning Initiative. Retrieved from https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md

Acknowledgements

The authors would like to acknowledge the helpful advice of Professor Robin Boast and Dr Kirsty Kitto, and the thorough support of both Ian Dolphin of the Apereo Foundation and Michael Webb from Jisc.