LACE

Learning Analytics Community Exchange

Learning Analytics Data Sharing Examples

A selection of brief case studies

By: Adam Cooper, Cetis

Published: 05 May 2015

Keywords: learning analytics, data sharing, feasibility, case studies

Sharing Learning Analytics data between organisations may sometimes be a highly desirable action, both for furthering research and as part of an expertise or technology based service. This document describes a number of examples to show how this can be the case, to illustrate similarities and differences, and to tentatively identify some of the possible causes of success or failure of sharing data for learning analytics.

Contents

Glossary of Abbreviations

API – Application Programming Interface, the means by which software components exchange data or direct processing.

CSV – Comma Separated Values (also generalised to Character Separated Values), a simple textual representation of tabular data.

ITS – Intelligent Tutoring System.

LMS – Learning Management System.

MOOC – Massive Open Online Course, commonly nuanced to xMOOC or cMOOC according to whether a conventional instructional style or a connectivist style of course organisation is used.

REST – REpresentational State Transfer, an architectural style for APIs that exploits the architecture of the web.

SaaS – Software as a Service, a kind of Cloud Computing.

VLE – Virtual Learning Environment, a regional term approximately equivalent to LMS.

XML – eXtensible Markup Language, a textual representation of structured information.

Introduction

Many applications of learning analytics require large scale data for educational data mining techniques or multi-variate statistics. Although the data from an institutional learning platform or a MOOC may be considered large, the scale and coverage of such datasets may be insufficient to allow the potential of learning analytics to be fully realised because of the great variety of learner and contextual attributes. This challenge applies to both learning science research and to potential products and services built around data generated during learning activities[1], and motivates the idea that data sharing between organisations - potentially including public and private sector bodies - could be an important enabler for effective learning analytics.

Data sharing is also indicated by models of service and IT provision, where expertise or technology is provided by a separate organisation to the education provider.

Questions about feasibility and sustainability remain, however, from socio-cultural, technical, and business perspectives. At present, in late 2014, it is appropriate to consider what examples can be found to help to point to answers to these questions, examples that illustrate the scope of successes, causes of failure, and possible tactics. In this respect, we do not understand the idea of "learning analytics data sharing" to represent a single idea, or a single idealised "solution", but to be an umbrella term for a diverse, but as yet unknown, range of different approaches that each navigates a different set of contextual factors.

It is the purpose of this document, which is currently a work in progress, to draw out some of these examples. It forms part work by the LACE project - comprising of a series of publications, face-to-face workshops, and online discussions - aimed at stimulating a broad range of stakeholders to make sense of what kinds of data sharing platform may be feasible in different settings, and to clarify where further conceptual or technical work may be needed. At present, many of the examples come from North America and the LACE project team would be very interested to hear of examples from elsewhere in the world, particularly Asia, South America, and Europe, because we believe that seeing examples from a range of different socio-cultural situations will help to illuminate the question of feasibility for people in all regions of the world.

To complete this introduction, it is appropriate to indicate the variety of data that might be involved and clarify what we mean by "data sharing", or "data sharing platform", in this document and in the LACE project[2].

The variety of data that is relevant to Learning Analytics is potentially very great indeed. For the purpose of this set of case studies, our interest is primarily in data about people or their activity in a learning-related situation. This excludes national or international classification schemes for subject

[1] For example, the activity of learners on a single course is likely to be so diverse that a learning resource recommender system would be practically useless if only based on data at this scale. This issue is discussed in: Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval, E. (2011). Dataset-driven research for improving recommender systems for learning. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 44–53). New York, New York, USA: ACM Press. doi:10.1145/2090116.2090122

[2] The work is being undertaken as a task in the work package dealing with interoperability and data sharing, where our over-arching view is that technical, semantic (i.e. concepts and meaning), and organisational factors must be taken into account.

matter of courses or learning resources, for example. These have been ruled out since, once the scheme is agreed, sharing such reference data is a trivial matter, so the topic is not interesting in a discussion of data sharing. Even within our chosen scope, questions about privacy, feasibility, sustainability, etc vary greatly depending on whether the data in question is about records of education achievement and history, public social media activity, LMS/VLE activity, etc. There is, therefore, unlikely to be a single approach to data sharing.

"Data sharing" is being used to refer to data sharing between legal entitles (an organisation or person), and not just sharing between software under control of a single legal entity. This outlook on data sharing requires careful consideration of many more factors, so when we use the term "data sharing platform", we are not using "platform" as a software engineer would tend to; a learning analytics data sharing platform should be a rich structure comprising technical, operational, business/funding, policy, and governance factors.

To summarise the aim of this document: it is intended to describe pertinent examples, and discuss issues, rather than to draw conclusions about what is feasible and viable in the future.

Structure of the Case Studies

This section lists and describes the aspects that will be considered in each of the chosen examples of data sharing. The examples appear in the following section.

Aims and Context

What is the educationally-relevant purpose of the activity, in terms of the kind of knowledge that is gained from the data, or the kind of actions that will be taken as a consequence of the analytics[3]? What breadth of learner contexts is included, including phase of education, geography, ...?

Data

What kinds of learner data are shared? This aspect is concerned both with the kind of activity or learner attribute is to be analysed, and also the nature of derived data and results of the analytical process.

Parties

What kinds of party participate in a sharing relationship, and what asymmetries exist?

Motivation

What motivates the sharing, what specific benefits is data sharing believed to deliver that would be difficult or impossible to achieve within a single educational establishment? What is the business or funding model that is founded on this motivation?

Non-technical Platform

What kind of policies, procedures, and contractual relationships exist?

[3] The motivation for sharing is considered separately, see below.

Technical Platform

What kind of technical approach is used? This could include choice of core technology, cross-party architecture, data exchange formats, and the role for standards[4]?

The Examples

Small case studies are presented for each example, following the structure outlined above. Web addresses for further information are provided at the end of each section.

Pittsburgh Science of Learning (PSLC) DataShop

Aims and Context

The PSLC DataShop is a repository and collection of analysis and visualisation tools to support a research community in the field of intelligent tutoring systems (ITS), in particular those systems based on the idea of Knowledge Components, which have their heritage in cognitive models of psychology. Knowledge components are defined on the PSLC wiki[5]:

> *"A knowledge component is a description of a mental structure or process that a learner uses, alone or in combination with other knowledge components, to accomplish steps in a task or a problem."*

Its immediate aim is to support the continued development of ITS, but it should be noted that much of the data in DataShop comes from systems that used by many thousands of school students, in grades 6-12, in the United States of America.

Data

The data input to DataShop consists of ITS log data, which comprises both the learner activity and the "tutor" (software) response. Learner logging captures responses to problems but can go down to the level of interaction with individual user interface components, and tutor logging includes presumed attainment of Knowledge Components.

Data export fields are essentially the same as imported fields, but with the potential for roll-up (e.g. individual interactions rolled up to student-problem level) and associated computation of simple aggregate statistics.

DataShop visualisation and reporting tools allow for general purpose manipulation and visualisation of attributes such as error rate and step durations at various levels of aggregation. They also allow for visualisation of learning curves, i.e. displays of the changes in student performance over time. These are based on the Knowledge Component paradigm.

Parties

DataShop is operated by researchers at Carnegie Mellon University (CMU) and is intended to serve researchers undertaking formal research studies. It is possible for researchers to use the DataShop

[4] The term "standard" is used loosely here to include any specified data structure agreed for use by several parties, hence we include open standards but are not presuming open standards.
[5] http://www.learnlab.org/research/wiki/index.php/Knowledge_component

as a private repository, with total control of who else may access the data they deposit. Depositors may make data-sets openly accessible, or may allow access requests. CMU appears to have no privileged position with respect to data access.

Motivation

The motivation for sharing is the more efficient development of knowledge by enabling multiple research enquiries to be based on a single data-collection activity. DataShop was established using grant funding from the US National Science Foundation. It is understood that, at present, it can be maintained with modest levels of funding contributed from research projects making use of the DataShop.

Non-technical Platform

The default presumption is that data-sets added to DataShop remain private to the depositor. Since DataShop is operated by, and for, a research community, primarily in the USA, the policies in place assume and require adherence to ethical research practice, including review by institutional ethics committees (IRB, Institutional Review Boards, in US terminology). DataShop reviews evidence provided by depositors that the data has been ethically collected and appropriate consents obtained, before permitting depositors to share it with select other researchers, or publicly. All data must be de-identified.

Technical Platform

DataShop is a centralised repository with import via XML and tab separated (CSV) files, and export of tab-delimited data. Importing data is a batch process, in which the uploaded XML/CSV files are submitted to a queue. The XML format, which is extensively documented as a public specification, the Tutor Messaging Format, provides for a richer data import. This specification has evolved over time, but remains specific to PSLC.

Further Information

The DataShop repository home page provides a listing of public data-sets:
https://pslcdatashop.web.cmu.edu/

The “help” page contains information on the technicalities, procedures, sharing policies, and reporting features of the DataShop:
http://pslcdatashop.web.cmu.edu/help.

This information is summarised in a “cheat sheet”:
http://pslcdatashop.org/downloads/DataShopCheatSheet.pdf.

The Tutor Messaging Format is an XML specification for logging data:
http://pslcdatashop.web.cmu.edu/dtd/guide/tutor_message_dtd_guide_v4.pdf.

Predictive Analytics Reporting Framework

Aims and Context

The Predictive Analytics Reporting (PAR) Framework focuses on Higher Education student retention and completion, across a range of types of institution in the USA (two and four year courses of study, public and private, traditional and non-traditional institutions). It undertakes benchmarking, prediction, and work to understand the signs of risk vs progress to completion. In addition to prediction, the aim is to support the identification of good practice in student retention through data analysis, shared models, and benchmarking across institutions.

Data

The PAR Framework uses a fairly small number of variables compared to all of the data that could be collected in a learning analytics scenario, although over 60 are collected. These come under headings such as demographics, course of study information, academic records, and institutional type/approach.

Student-level predictions of risk, and aggregated benchmark data is returned to member institutions.

Parties

The PAR Framework is managed by the WICHE Cooperative for Educational Technologies (WCET), a non-profit organisation. They describe PAR Framework as a collaborative, which currently consists of 16 institutions. Each of these institutions contributes its data to a central database managed by WCET, and receives the results of student-level analysis on its own data; WCET maintains a team including data scientists. Benchmark data is available to all member institutions.

WCET appears to not make its own use of the data, but to be a centre of expertise and provider of technology to the members.

Motivation

The PAR Framework motivations are two-fold: a) that there is a cost-saving in having a central analytics service with highly skilled staff, covering multiple aspects of expertise from data science to policy and HE practice; b) cross-institutional benchmark studies provide valuable information on effective strategies to promote achievement, engagement, and progress.

Initial setup funding, and funding to support the initial 16 institutions was provided by the Bill and Melinda Gates Foundation. The intention is that member institutional partners will provide sustainability through subscription fees and the PAR Framework is intended[6] to become an independent legal entity from 9 December 2014.

Non-technical Platform

The PAR approach is build around a collaborative/cooperative model and governance is member-led. Each member institution is required to follow its normal institutional approval process for human

[6] See http://www.prweb.com/releases/PAR/Independence/prweb11888897.htm

subject research (ethics committee/IRB) and the PAR team all have certification in human subject research. PAR indicates attention to data privacy, record security, and research integrity.

Technical Platform

The technical platform is essentially invisible; WCET clearly operates an enterprise database and appears to accept data in batch files.

The PAR Framework publishes its set of common data definitions under a Creative Commons licence. These are not fully-specified, and are not mapped to data definitions from other sources.

Further Information

Information about PAR and their approach:
http://wcet.wiche.edu/par/about

Data definitions:
https://community.datacookbook.com/public/institutions/par

Open Academic Analytics Initiative

Aims and Context

The Open Academic Analytics Initiative (OAAI) is a completed project, largely undertaken in 2012 involving two US community colleges and one Historically Black College and University in addition to the lead, Marist College, a private US higher education institution. It focussed on predictive modelling of students at risk of dropping out with a view to directing interventions, and researching different intervention strategies.

The project is of particular interest because it considered the portability of predictive models, i.e. model sharing, both technical aspects and statistical performance.

Data

Just short of 30 data elements were used, of similar type to those used in the PAR Framework, largely relating to demographic details and academic performance, but with addition of grade-book and tool usage counts from Sakai.

Parties

This was a project with no ongoing data sharing.

Motivation

The particular motivation for data sharing was the investigation into the portability of predictive models across diverse academic contexts (but all US post-secondary).

Non-technical Platform

This was undertaken as human subject research; ethics committee (IRB) approval was gained in all institutions.

Technical Platform

OAAI is distinctive in that the project specifically set out to develop and deploy a system comprising a number of Open Source Software components. The principal point of interest for this case study is their use of an open standard, PMML (Predictive Modeling Markup Language), to export their predictive models.

Further Information

The OAAI project page outlines the project and its results, with abundant links to reports and technical documentation, as well as videos:
https://confluence.sakaiproject.org/pages/viewpage.action?pageId=75671025

A description of the data elements used:
https://confluence.sakaiproject.org/download/attachments/75671025/Required%2BDataset%2Bformat.docx

MUltimodal contextualized Learner Corpus Exchange (MULCE)

MULCE was a research project supported by the National Research Agency in France between 2007 and 2010, predating all of the other examples in this document. The MULCE team worked on requirements for research data to be shareable from a variety of perspectives and built a repository. It is now hosted by the LRL (Laboratory for Research on Language) in the MSH (Maison des Sciences de l'Homme) in Clermont-Ferrand and used for research on multi-modal interactions in language learning. “MULCE” will be used in the present tense to refer to the project and to continuing activity together. Since MULCE has operated for several years, albeit at a modest scale, it has learned some of the lessons of data sharing for a research community (see Further Details) including the practicalities of exchanging and re-using copora, structuring multi-modal data and contextual information, and ease of access for deposit or by analytical tools.

Aims and Context

MULCE seeks to support an improvement in the quality of academic research in technology enhanced learning by broadening access to data-sets.

Data

There are two distinctive features of the MULCE repository: the variety of modalities of interaction, including virtual worlds and white boards as well as synchronous and asynchronous text, for example; a distinction between “global” datasets, which are the original form, and “distinguished” datasets which are the selected and transformed data used in particular research. MULCE is also concerned with capturing, in variously structured forms, metadata to describe the context from which the learner activity data was captured.

Parties

MULCE is, in principle open to any interested researcher to deposit data, although it remains necessary for MULCE team members to be involved in the process for technical and quality assurance reasons. The repository is open access for data download. The terms of use specify non-commercial use and note that the intended users are researchers and teachers.

Motivation

What motivates the sharing, what specific benefits is data sharing believed to deliver that would be difficult or impossible to achieve within a single educational establishment? What is the business or funding model that is founded on this motivation?

Non-technical Platform

The standard licence for use of data in the MULCE repository is Creative Commons Attribution Non-Commercial Share-Alike, although there are clauses which permit a depositor to add conditions. The terms of use indicate, but do not require (the word "should" is used), that users should alert the depositor if personal data is discovered.

MULCE has clear expectations for the level of detail of the context of the data, including what might be called the learning design as well as the higher level aims of the course.

Manual intervention from the MULCE team forms part of the deposit process.

Technical Platform

MULCE has taken a mixed approach to data exchange formats, preferring to use or be inspired by published interoperability standards while developing its own XML-based structures[7] for expressing the activity and contextual information. They use, or refer to, for example: the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), IMS Content Packaging, IMS Learning Design, and IEEE Learning Object Metadata (LOM). OAI-PMH permits the harvesting of the metadata about resources (in this case datasets) so that they can be indexed by, and discovered through, other services. i.e. they embrace the idea of a distributed architecture for resource management and discovery.

MULCE also developed software to support the de-identification of learners in text-based chat and forum exchanges. Source code is available, but it dates from 2006.

Further Information

Reffay, C., Betbeder, M.L. & Chanier, T., 2012. Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project. International Journal of Technology Enhanced Learning, 4(1/2), p.11. Available at: *https://edutice.archives-ouvertes.fr/edutice-00718392/document* [accessed 2015-05-06].

MULCE main site and documentation (English and French site):
http://mulce.org

Browsable MULCE Repository with English and French versions:
http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/VIEW/PUBLIC/01/accueil.jsp

[7] More recent work associated with MULCE has explored the idea of deposit in more idiosyncratic representations along with the use of a shared ontology which is mapped to it for research. Chebil, H., Courtin, C. & Girardot, J.-J., 2012. The Proxy Model: A New Approach to Sharing and Analyzing Learning Traces Corpora. International Journal of Information and Education Technology, 2(4), pp.308–311. Available at: http://www.ijiet.org/show-32-107-1.html.

Stanford Data Portal for Research

This is also sometimes known as the Stanford VPTL (Vice Provost for Teaching and Learning) Data Portal. The term "VPOL (Vice Provost Office for Online Learning) Learning Data" is also used, apparently interchangeably.

Aims and Context

The Data Portal is intended to be used for academic research without any presumption of research topic, within the bounds of what is possible with the data.

Data

Data from MOOCs operated by Stanford University is available, specifically from the use of three platforms: NovoEd, Coursera, and OpenEdX. The data describes assessment, forum, and video-watching activity. Interactions are available at an intermediate level of tracking (e.g. assessment responses and excerpts from video track-logs). A computed time-on-task is also available.

Parties

The data is available for academic research by scholars from Stanford and elsewhere. NovoEd data is only available to Stanford.

Motivation

The motivation is presumed to be a combination of Stanford University seeking to enhance its reputation and to manage exploitation of the data by its own academic staff ("faculty"). It appears to be entirely self-funded by Stanford, a well-endowed institution, so likely to persist in the medium term.

Non-technical Platform

Stanford's MOOCs make it clear that data from learner activity will be used for research as part of its Terms of Service, which were revised following the 2014 Asilomar Convention (see Further Information). Access to the data requires assent to a data use agreement and formal assessment of all requests for data access by the VPTL Data Sharing Working Group, which reserves the right to revoke its consent. Requests are required to specify the project for which the data will be used and to limit use to that purpose. Disclosure to third parties is explicitly prohibited, as is publication of any personally identifiable information. Furthermore, data users are prohibited from attempting to re-identify individuals.

Technical Platform

The data is held in a MySQL database with a stanford.edu URL. Notes are available explaining access from Excel and SPSS. For OpenEdX and Coursera course data, the structure of the database is a replica of that used in the production platform. MOOCDb versions are also available for some courses; MOOCDb is a project founded at MIT to develop a platform-agnostic relational data model for MOOC data.

Further Information

VPTL Data Portal Home Page:
http://vpol.stanford.edu/research

Data use agreement:
https://stanford.box.com/datauseagreement

Description of data structures:
http://datastage.stanford.edu/

Asilmar Convention, six principles agreed on by a meeting of scholars concerned with "the collection, storage, distribution and analysis of data derived from human engagement with learning resources":
http://asilomar-highered.info/

InBloom

Aims and Context

InBloom was a US non-profit with wide-ranging intentions to support progress tracking and delivery of individualised learning in schools. It wound up its operations in 2014, following widespread criticism and threats of legal action arising from the range of highly personal data that had been ingested into the InBloom databases.

Data

Since the InBloom ambition was a wide-ranging one, to support detailed progress tracking and individualised interventions by teachers, a very wide range of data fields were supported. Over 400 fields were supported, although it was a school district decision which fields were used, essentially the set of demographic, attendance, special needs, academic performance, and other data kept in school student records systems and reported to local authorities.

Parties

InBloom was essentially a technology provider to public education authorities, with each having a separate data store; to quote from their FAQ:

"States and districts are responsible for protecting student information. In accordance with both law and inBloom policies, inBloom is a third party service provider covered by FERPA, responsible for being an information steward to these public officials. InBloom provides each participating school district with its own protected storage space, so that district can continue to manage and control access to its own data, just as it always has."

Motivation

The data sharing that InBloom enabled permitted the development of learning-related software by third parties, with the software having access to data that it would not normally have.

Non-technical Platform

The centre-piece of the InBloom approach was to provide the technical platform and for the school districts to be in control of the data. InBloom assumed responsibility for data security, including monitoring procedures, and designed the software to allow school districts to control data disclosure to third party providers of learning-related software.

Technical Platform

The InBloom platform was essentially a SaaS solution with support for bulk data ingestion of XML files as well as extensive REST APIs to allow third party developers to build new software applications.

The data definitions were adopted from the Common Education Data Standards of the US Department of Education.

Further Information

The InBloom website appears to be no longer available, but snapshots from before the announcement of their intention to wind down may be found in the Internet Archive "Wayback Machine": http://web.archive.org/web/20140301131936/https://www.inbloom.org/.

There are numerous articles discussing the closure of InBloom, for example: http://www.slate.com/blogs/future_tense/2014/04/24/what_the_failure_of_inbloom_means_for_the_student_data_industry.html.

The Common Education Data Standards from the US Department of Education: https://ceds.ed.gov/.

xMOOC Platforms

xMOOC platforms are better described as shared delivery platforms, rather than shared data platforms, but they are included for three reasons:

- these platforms typically offer the usage data back to their clients;
- the scale and relative uniformity and simplicity of the delivery model[8] makes the data attractive for learning analytics;
- MOOC data is generally very difficult to fully anonymise since forum posts often contain either direct identification of the sender, or sufficient traces to allow for identity to be inferred when combined with social media and other public-access data.

The initial round of xMOOC platform implementation focussed on the delivery platform and the provision of database dumps to their client/partner, while apparently making little central use of the data. Although little evidence of central use of the data is available, the current privacy policies are rather similar to Silicon Valley social media corporations, and markedly different from PSLC DataShop or the PAR Framework. Coursera usage data may, for example, according to their privacy policy, be shared with any business partner for research or to allow them to "share information about their products and services that may be of interest [to you]." Futurelearn includes statements

[8] The vast majority of course use a combination of video lectures, short assessment quizes, and online forums, while many also include mid-point and final graded assignments, sometimes deploying peer assessment.

with a similar sentiment and a similarly open-ended clause that "your information may be used by us and by technology partners and course and content providers chosen by us."

The xMOOC providers' initial focus on delivery platform rather than data use is now beginning to change, doubtless for a variety of reasons including: an ability to attend to less mission-critical ideas, now the core platforms are stable; a need to evolve sustainable business models; a need to compete on demonstrable effectiveness of the platform; an increasing interest from clients as they move from initial adoption into a more reflective stage of delivering MOOCs. This change is reflected in, for example, Coursera recruiting a Director of Analytics[9] with a role including both platform-level analytics (e.g. A/B testing) and services to partner institutions ("university partners need the data insights we can provide them to advance their pedagogy").

The European Multiple MOOC Aggregator (EMMA), a multi-lingual platform, includes learning analytics in its scope but is still developing its offering.

Further Information

Coursera privacy policy:
https://www.coursera.org/about/privacy.

Futurelearn privacy policy:
https://about.futurelearn.com/terms/privacy-policy/.

The European Multiple MOOC Aggregator (EMMA):
http://europeanmoocs.eu/blog/about/.

Open Data Initiatives

This section does not describe a single example of a learning analytics data sharing as scoped-out in the introduction, but considers two initiatives at the extreme end of data sharing: open data. LinkedUp is a project specifically targeting web data for education, which "aims to push forward the exploitation of the vast amounts of public, open data available on the Web, in particular by educational institutions and organizations". The second initiative is the Open Knowledge Foundation (OKFN), which is also a member of the LinkedUp Project.

OKFN is concerned with furthering access to, and use of, knowledge in its widest sense, and through a number of means including campaigning, community support, consulting, and software development. It promotes publication of public sector spending data on the one hand, but also operates an Open Education Working group, about which it says "Open Education is much more than just OER and involves aspects like opening up relevant educational data and changing both institutional and wider culture". In this case, and at present, however, "relevant educational data" is not thought of as the kind outlined in the introduction: data about people and their actions.

Putting to one side any speculation about what the Open Education Working Group may do in the future in the intersection between Open Education and Learning Analytics, the most significant contribution, from OKFN, of relevance to a discussion of learning analytics data sharing platforms is

[9] https://www.coursera.org/about/careers/9d2e3d7a-e391-4197-bd5e-f2ed107bc800

their open source data portal platform, CKAN, the Comprehensive Knowledge Archive Network. CKAN is used by numerous city and national government open data publication programmes.

CKAN has a number of features that may make it relevant to some realisations of learning analytics data sharing:

- Dataset metadata can be harvested, allowing for search across a federation of CKAN instances while data may be held back with access subject to separate agreement.
- Aside from being open source software, the CKAN architecture provides for "extensions" to be written without requiring source code changes in the main CKAN software. There is, for example, an extension to accommodate professional-level geospatial data in CKAN.

The LinkedUp project is responsible for the Linked Education Cloud catalogue of open data on the CKAN-powered datahub.io site. At present, this includes a small number of datasets of relevance to learning analytics, but outside the scope given in the introduction to this paper, for example machine readable catalogues of learning objectives from the Achievement Standards Network. It includes only one dataset within scope, although aggregated and of tangential interest for learning analytics, which is circulation data from the University of Huddersfield library.

Further Information

The Linked-Up Project:
http://linkedup-project.eu/

OKFN mission and methods:
https://okfn.org/about/

CKAN open source software for data publishing:
http://ckan.org/

Linked Education Cloud:
http://datahub.io/group/linked-education

Discussion and Open Questions

While it remains impossible to draw any strong conclusions from the small number of data sharing initiatives within the scope set out in the introduction, it is possible to point to some features of interest. The examples also invite some open questions, a number of which are posed to draw this document to a close, consistent with its intended purpose of stimulating debate on the topic of learning analytics data sharing.

The DataShop and PAR Framework examples are, arguably, initiatives with a good chance of sustainable activity, although both have been initiated with grant funding and neither has yet demonstrated longevity. Their good chance of sustainability rests of a number of factors, which they have in common, in spite of serving quite different kinds of communities:

- Ethical and privacy issues have been dealt with within the framework of human subject research practice.

- They are purposeful with respect to scope, with DataShop focussing on a particular research field (an approach to cognitive modeling and ITS) and PAR on the problems of retention and completion.

They also differ in a number of respects, in addition to their purpose:

- Whereas DataShop works with anonymised data (and with data that is plausibly anonymisable in practice), PAR requires that the results of analysis can be matched to individuals.
- DataShop is concerned with school-age subjects, whereas PAR with post-compulsory education students.

OAAI is interesting in that they investigated sharing of models, rather than student-level data. This naturally avoids most concerns about breaches of privacy[10], even if these models were published as open data. While it is more likely that the management of educational establishments would be more likely to entertain model sharing within a closed group, rather than open publication, the research community's move towards open scholarship and reproducible research indicates a role for model sharing as open data.

The failure of InBloom has been much talked about, and yet they appear to have done nothing evil with learner data, and suffered no breach of data security. This suggests that its downfall was fundamentally caused by a failure of trust.

xMOOC platform providers have moved very quickly to provide a service that we did not anticipate only a few years ago. It is plausible that they will make increasing use of the data that they collect but their attitude with respect to data ownership and disclosure suggests that sharing will be in the mould of social media providers, and driven by their business interests. It remains to be seen whether the future sentiment of xMOOC learners and course providers turns against this kind of approach, and whether projects such as EMMA develop a sustainable alternative. It also seems likely that participants in the OKFN Open Education Working Group, and innovators in the cMOOC ecosystem, may have quite different visions of future in the intersection between Open Education and Learning Analytics than xMOOC providers.

Open questions include:

- With respect to the norms of human subjects research, how well matched are they to "Big Data" approaches and to aspirations for learning analytics?
- If we assume that the failure of InBloom was essentially a failure of trust, what could be done to either build trust or to temper ambition to require less trust?
- Is model sharing a viable approach? Particularly for operational use, will more statistically-powerful models necessarily embed local contextual information?
- Do approaches exist for sharing data that cannot be properly anonymised (other than cases similar to PAR, where the student-level data is shared only with a trusted data processor)?

[10] The risk remains when models are built using small numbers of students, and with poor choice of attributes.

- Would it be useful to share access to data, while keeping the individual records private? For example, to allow a wide range of statistical analyses using an arbitrary set of attributes to be undertaken as a “black box” activity.
- At what point in time, and for what kind of data, will it become sensible to seek to agree standardised data definitions?
- What are the future prospects for LA data sharing by xMOOC platform providers, and how quickly will change happen?

Additional Material

LACE Project blog posts on the topic of privacy may be found at: *http://www.laceproject.eu/blog/tag/privacy/*. These include articles on privacy by design and the role of systems of trust. The posts are open for comment.

The home page of the 1st Learning Analytics Data Sharing Workshop, held at EC-TEL 2014, is at: *http://www.laceproject.eu/lads14*.

This document formed the first part of work on feasibility and roadmapping of data sharing for learning analytics, which is reported on as Deliverable 7.2, which is available at: *http://www.laceproject.eu/deliverables/*

About ...

Acknowledgements

The author would like to thank Tore Hoel and Hendrik Drachsler for helpful comments.

This document was produced with funding from the European Commission Seventh Framework Programme as part of the LACE Project, grant number 619424.

About the Authors

Adam works for Cetis, the Centre for Educational Technology and Interoperability Standards, at the University of Bolton, UK. He rather enjoys data wrangling and hacking about with R. He is a member of the UK Government Open Standards Board, and a member of the Information Standards Board for Education, Skills and Children's Services, and is a strong advocate of open standards and open system architecture. Adam is leading the work package on interoperability and data sharing.

About this document

For more information, see the LACE Publication Policy: *http://www.laceproject.eu/publication-policy/*. Note, in particular, that some images used in LACE publications may not be freely re-used.

The permanent location for the latest version is: *http://www.laceproject.eu*/publications/public-drafts/wp7-dse.pdf

About LACE

The LACE project brings together existing key European players in the field of learning analytics & educational data mining who are committed to build communities of practice and share emerging best practice in order to make progress towards four objectives.

> *Objective 1 – Promote knowledge creation and exchange*
> *Objective 2 – Increase the evidence base*
> *Objective 3 – Contribute to the definition of future directions*
> *Objective 4 – Build consensus on interoperability and data sharing*

http://www.laceproject.eu *@laceproject*