

# Learning Analytics Interoperability – some thoughts on a “way ahead” to get results sometime soon

*Adam Cooper, Cetus, October 2013, Draft v0.2 (subject to substantial refinement/rework)*

## 1 Why Did I Write This, What is Implied by the Title?

This document has been written to express my thoughts on the resolution of tension between several factors pertaining to interoperability in relation to learning analytics. These factors are:

- The difficulty in getting standardised data out of information systems in a consistent way is a barrier to conducting learning analytics. There is a need now.
- There is a hunger to taste the perceived benefits of using learning analytics.
- The scope of data relevant to learning analytics is enormous. To reach the minimal common ground necessary to declare “a standard” or interoperability across all of these is intractable given available human resource because experience shows either analysing the breadth of actual practice or defining anything by consensus is slow.
- The range of methods and targets of learning analytics is diverse and emerging as experience grows. This places limits on what it is rational to attempt to standardise. In other words, we don't really know what LA is yet and this brings the risk that any spec work may fail to define the right things.

The “results” of the title are the situation where increased interoperability contributes to practical learning analytics (exploratory, experimental, or operational). The way ahead to get results sometime soon requires care; focussing on the need and the hunger without restraining ambition will surely mean a failure to be timely, successful, or both. On the other hand, although it would be best (in an ideal world) to spend a good deal of time characterising the scope of data and charting the range of methods and targets, it is feared that this would totally block progress. Hence a middle way seems necessary, in which a little time is spent on discussing the most promising and the best-understood targets. i.e. to look for the low hanging fruit. This represents a middle way between the tendencies of the sales-man and the academic.

A documented problem definition is an essential initial step for several reasons:

1. it provides a means to communicate intent with people who might join the effort, or who have a view on priorities or feasibility such that this vital input can reduce false start risk;
2. it is very helpful for new joiners to a collaborative effort;
3. it gives collaborators a better chance of being "on the same page";
4. it is logical to define the problem before the solution;
5. it helps guide research into prior work, to allow relevant input technology (in a general sense) to be discriminated from the irrelevant;
6. properly articulated written documentation flushes out inconsistent thinking, locates blind-spots, etc better than ad-hoc oral discussion.

## 2 Background Perspectives

### 2.1 Data Consumers and Implications

The rest of this document is written with two groupings of learning analytics scenario, grouped according to the kind of software, in mind:

- Informative visualisations potentially embedded in an existing product (maybe as dashboard widgets), or self-contained applications. These may be for use of learners, teachers, or management.
- Analytical tools ranging from Excel to Tableau, R and SPSS etc

Across this range of consumers, different data-bindings make for low-friction use. Hierarchical XML is probably not the way to go; flattened modeling, “denormalised” representations, and binding to JSON or CSV probably is. JSON-LD may introduce obstacles. XML, or at least the data-containing part, should be isomorphic to a table if XML is used. Acid test 1: how many lines of code (excl existing libraries), or manual operations, does it take to fetch data and show it on some kind of chart? Acid test 2: is the data structure amenable to efficient streamed processing (or splitting in Hadoop etc)?

Aim for structural uniformity across a wide range of source data and problems in focus. A tabulation achieves this by avoiding pollution from syntactic or conceptual idiosyncrasies that is likely when a structure is framed in a particular technology or modeling formalism.

## 2.2 Spectrum of Data (exchanged)

*This looks at one dimension of the data and some implications that arise depending on where in the spectrum we elect to attempt interoperability. Roughly-speaking, from red to violet requires more statistical &/or algorithmic processing.*

Red – raw log – needs a lot of processing, is low on meaning, interpretation requires a lot of information about the context of capture – the recipient has lots of flexibility but has to do a lot of work and needs to “know” details of the data source to make sense. This analysis software may be operated by a data wrangler engaged in exploratory work or running on a periodic job and is likely to employ compute-intensive algorithms to generate statistics used downstream for user-access-time processing/visualization. [Implication: only very common or high level semantic units are defined, the spec is essentially syntactic + an API with accommodation for good use of vocabularies]

Green – descriptive stats from system at individual<sup>1</sup> level – mid-level meaning, the system is programmed to provide data we can understand but it is still hard to extract actionable insights without further processing – the recipient has quite a lot of flexibility and can take on many forms, some of which do not need to process the data further while others will.

Violet – the data consumer receives a prediction etc – meaning is clear and actionable – recipient has little scope for variety since it gets the answer to a predetermined question using a predetermined method.

## 2.3 Analytics for Information or Insight?

Getting data out of a system to show it as a bar/line/scatter/pie/whatever chart barely touches the surface of “analytics”. Getting insight is likely to require the application of statistical and machine learning methods and these may require data elements or treatment that would not be necessary for basic information presentation.

## 2.4 Scope of Education

In principle:

- online
- traditional bricks and mortar HE
- K12

## 2.5 Client/Subject/Object

When thinking about use cases, I think it is helpful to be clear about different configurations of the analytics object, subject and client because it is easy to conflate them<sup>2</sup>.

“Analysis object” is the entity that will be acted-upon differently as a consequence of the use of analytics. This may often be the same as the “analysis subject” - the entity that the data is about – but need not be so. For example, data about student success (the data subject is a student) could be used to understand aspects of teacher

---

1 Probably a person, but maybe a content-resource or another entity instance. We might add “Blue-green – as green but aggregated to entity-type (class) level.”

2 The following paragraph was taken from my white paper on a framework of characteristics for analytics, available at <http://publications.cetis.ac.uk/2012/524>.

behaviour (the analysis object is a teacher), course design or even whether the course marketing is attracting students of the “right” kind. “Analysis clients” are those who use the results of applying analytics. Analysis subject, object and client may sometimes be identical, for example when analytics is used for self-regulation.

## 2.6 [nitty detail] Accommodating Different Markets

Define vocabularies/metrics to accommodate different localisations of essentially similar concepts, where “localisation” might be geographic or related to institution type or stage of education. For example, specialisations of what is essentially the same concept realised in different settings could be managed through a light-weight mechanism and software conformance declared at unlocalised + one or more localisations.

## 3 Promising Targets

This section provides some sketches of answers to the question: “what, from the perspective of an educational establishment, are the most pressing drivers for adoption of learning analytics?”

Simply being able to get data out to visualise in a different way to what the LMS provides, for example, is not really a compelling case for investment.

### 3.1 Retention & Engagement (generalises to early intervention)

#### 3.1.1 Characteristics of the Target and a Plausible Response

An important characteristic of this target is that it is primarily engagement that is measured and managed, rather than learning. This simplifies the analytical problem and reduces the risk of over-stepping the role of the educator, whose capacity to deal with subtleties exceeds our number-crunching.

The quality of source data is also highly likely to limit what can be achieved with confidence; mining fine-grained data does not guarantee meaningful results if the quantity and quality of data does not match the size of the parameter space. Achieving valid and reliable results is likely to be easier if a less intrinsically-messy target is chosen and if data is somewhat aggregated before processing.

Data sets indicated for informative or predictive scenarios:

- high level attainment on course (existing specs exist)
- simplified measures of activity in LMS
- attendance
- library use

Additional sets of data only required in a predictive scenario:

- high level demographic (existing specs exist)
- prior performance (SATs, qual level etc)

Concrete examples where people have done this exist.

#### 3.1.2 Anecdotes to Support this Target

**On "Trak": First Steps in Learning Analytics**<sup>3</sup>, EDUCAUSE Review, 2012 (U. of Canterbury, NZ)

*“A current focus at the University of Canterbury is to provide an early intervention process to enhance student engagement, retention, and success.”*

**Report on a Survey of Analytics in Higher and Further Education (UK)**<sup>4</sup>, Cetus, 2013 (small UK survey)

Respondents identified as most significant drivers: attainment, retention and assessment; student satisfaction.

---

3 <http://www.educause.edu/ero/article/trak-first-steps-learning-analytics>

4 <http://blogs.cetus.ac.uk/adam/2013/09/16/report-on-a-survey-of-analytics-in-higher-and-further-education-uk/>

**Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations**<sup>5</sup>, ECAR, 2012 (US survey)

*“From these data, we can glean that the likely early benefits of analytics will be in the areas of student performance, student recruitment and retention, and resource optimization (but not extending to cost reduction).”*

There are many items in the EDUCAUSE library tagged under “Student Retention” and “Learning Analytics”<sup>6</sup>.

The very wide (world-wide) interest in the achievements, and approach, of the Signals Project at Purdue is significant<sup>7</sup>. Their approach is one of enhancing success by addressing student persistence, engagement, and integration into the institution rather than by focussing on what might, or might not, have been learned.

## 3.2 Assessment Response Analytics

### 3.2.1 Characteristics of the Target and a Plausible Response

Considering assessment taken in both online and offline modes but for which there is data relating to responses, not simply the marks/grade. The object of analytics might be from among: learner<sup>8</sup>, assessment, teaching method, learning environment, etc. The aim might be diagnostic, predictive, evaluative, or pertaining to understanding of the mechanism between intervention and effect, etc. The problem is moving assessment interactions, and optionally granular outcomes or with information on the learning objectives being assessed, into a statistical analysis/visualisation system or expert analyst process. This is NOT a gradebook and may, for example, analyse question quality, answering pathways, model learner knowledge development... The essential point is that assessment data only within the software that handles the assessment delivery or marking represents a huge missed opportunity.

Although analytics offers the potential to make a wider range of activities assessable from the activity logs they produce, or from data that can be captured by video monitoring etc, these are best left as outstanding research topics. Stick to assignments and tasks that are fairly commonly assessed and traditional in character; for a v1, it would also be sensible to de-scope assessment of collaborative tasks or participation in discussions etc.

Data sets indicated for informative or predictive scenarios:

- response data as IMS QTI results (which are not only applicable to QTI formatted assessments)
- potentially data about an objective that was being assessed (relationship between objectives kept out of scope)

The work to be done would essentially be of coordinating implementers and getting others on-board to exercise (and potentially revise) a part of QTI that has not yet received much attention (but which is based on practice in high stakes assessment organisations) with the added twist that there is more to assessment response analytics than assessment results reporting. This represents an obvious next step for adopters of QTI ASI as well as a sensible extension for e-marking vendors, LMSs etc (there is NO dependency on assessment content being in QTI format).

### 3.2.2 Anecdotes to Support this Target

A report in a piece entitled **Assessment Analytics**<sup>9</sup> describes recent activity in the UK that has been supported by the Joint Information Systems Committee (Jisc) under its Assessment and Feedback Programme and a Cetus Case Study **Acting on Assessment Analytics**<sup>10</sup> describes how Huddersfield University have used assessment analytics to help students modify behaviour based on e-marking data.

The topic is clearly seen as important by Blackboard<sup>11</sup>, although my we would probably not wish to emulate “... tightly integrated on the tool your faculty and staff know best: Blackboard Course Delivery.”

5 <http://www.educause.edu/library/resources/2012-ecar-study-analytics-higher-education>

6 <http://is.gd/aEuDKO> (38 items on 2013-09-30)

7 [http://www.itap.purdue.edu/learning/docs/research/Arnold\\_Pistilli-Purdue\\_University\\_Course\\_Signals-2012.pdf](http://www.itap.purdue.edu/learning/docs/research/Arnold_Pistilli-Purdue_University_Course_Signals-2012.pdf)

8 While the learner is likely to be a common (obvious to all) object, hence points to a must-have set of use cases, the very similar data may be able to be put to many other uses with consequence for learning and learner satisfaction/sentiment.

9 <http://jiscdesignstudio.pbworks.com/w/page/67696603/Assessment%20analytics>

10 <http://publications.cetus.ac.uk/2013/750>

11 <http://www.blackboard.com/platforms/learn/products/blackboard-learn/assessment-accreditation-analytics.aspx>

*This section could be extended.*

## **4 Possible Targets**

These would probably fail the “pressing drivers” test but may be good targets for other reasons.

### **4.1 Institutional LMS Benchmark Analytics**

Most institutions must surely have a strategy process that involves consideration of the shape of institutional LMS and associated tool provision, alongside strategies to promote more effective use (which could be viewed in terms of either changing, or better supporting existing pedagogies). This strategic process should reasonably be informed by data on patterns of use and some form of benchmark referencing (either between institutions or between faculties/schools within them since discipline-related factors are relevant).

The question of what makes a good statistic for this purpose is less contentious than what constitutes a good statistic for learning (at least if we dispense with the convenient fiction of attainment). This does not mean it is trivial, and the damaging consequences of fixing on mis-judged benchmarks in decision-making remain.

There are several institutions that have taken an analytical approach to LMS use when deciding on major changes such as LMS replacement, although sometimes the analysis was done but not taken account of<sup>12</sup>. These would provide a basis to fairly quickly define a v1, although care would be needed to define terms carefully to pick out common concepts between platforms.

The interest in this target substantially depends on being able to access comparison data from other institutions. There would be a role here for a consortium organisation or organisations (global, regional, or national).

### **4.2 xMOOC Data**

A slightly different kind of target. The stakeholders are different and the idea is a little more distant from believable business cases for [almost all] current educational institutions. This would be contingent on gaining interest from [at least some of the] significant players such as coursera, edX, FutureLearn, but this is conceivable.

The volume of data available from a MOOC is large enough that questions can be answered that would be impossible to do so with an acceptable level of statistical significance with a traditional education cohort. Many HE institutions are explicit that understanding their courses, with a view to influencing design (including traditional provision), is a motivation for experimenting with xMOOCs. Few, if any, institutions will (except maybe at the very start) use a single platform. This poses a challenge that work on common models and interoperability could help address: how to consistently apply the same analytical methods (assumed to be programmatic statistics and data mining, whether scripted or operated “as a service”) across xMOOC data from several platforms. It is conceivable that, “how can we get our data out and quickly extract some value” will be one of the top questions when decisions about xMOOC platform are made.

Work towards this target would focus on the identification and documentation of common concepts across several platforms<sup>13</sup>. It would not worry that each platform would provide information outside this common core; we would not seek a situation where all analysis could be applied over multiple platforms, merely that some common and in-principle repeatable analyses could be made more routine. In practice, data is likely to be provided as MySQL or JSON (etc) dumps according to the platform technology.

Usage scenarios are important. For some uses, it may be preferred to keep the source data in its existing form because analysis will inevitably need to stray away from the common concepts, and it would be unhelpful to force technology-switching by requiring the interoperable part to be in a different technology. In this case, interoperability might be akin to defining a common front end of a database view, with a platform-specific mapping at the “back end”. Other uses may indicate ETL, but the essential character remains as being a transformation after data export from the platform. Platform compliance would be indicated by a clear mapping (and possibly, for example, ISO SQL). For the case of fine-grained data, realistically as JSON or text log-file...

---

<sup>12</sup> This paper from SFU tells an interesting, and I suspect not unique, story: [www.ifets.info/journals/15\\_3/11.pdf](http://www.ifets.info/journals/15_3/11.pdf)

<sup>13</sup> Necessarily including those deemed “MOOC platforms” but not excluding what would normally be labelled “LMS platforms”, for which there is existing work on simple common models (e.g. Anna Dyckhoff eLAT, RWTH Aachen).

that is probably best left alone (see “Spectrum of Data”).


## 5 *Tactics-now*

*The imperative is to do something useful with a good chance of adoption, recognising that we can neither provide a complete solution nor have the definitive view on the scope of “learning analytics”. A pragmatic course of action is likely to be one that gets people able to do the kind of thing they are doing now, but more efficiently or with greater agility, and which then enables the next few steps in their learning analytics journey<sup>14</sup>.*

In my view, near term tactics to get going would involve:

- Focus on one or two targets and describe scope in writing.
- Explicitly identify evidence of demand or be clear about the plausible business cases that would motivate educational organisations and suppliers to both engage in the standard development and adoption process.
- Narrow-down range of data, attend to the minimal viable dataset to realise value rather than visions of a beautiful solution.
- Build on knowledge of data use in practice rather than theorise about what might be.
- Borrow existing data structures if possible (domain-specific and generic statistics/data-mining<sup>15</sup>).
- Go “green” in the spectrum.

## 6 *IPR Declaration*

This document was created by Adam Cooper (Cetis, University of Bolton) and is licensed , <http://creativecommons.org/licenses/by/3.0/>.

As far as I know, it contains nothing that is the subject of a patent claim.

---

14 Plausibly, a journey from almost exclusively informative approaches towards greater use of analytics to support insight, where the data can support this.

15 e.g. SDMX, datacube are intended for data publishing. Although these structures are attractive in the degree to which statistical metadata is modeled, they present a significant processing overhead so it would probably be wise to NOT use them. Whereas they are intended to allow published statistics to be self-documenting with respect to the concepts, we would expect standardised definitions to be external to the data. ARFF is one step up from absolute minimum <http://weka.wikispaces.com/ARFF>