E-Learning Standards and Learning Analytics

Can Data Collection Be Improved by Using Standard Data Models?

Ángel del Blanco, Ángel Serrano, Manuel Freire, Iván Martínez-Ortiz, Baltasar Fernández-Manjón
Complutense University of Madrid
School of Computer Science, Department of Software Engineering and Artificial Intelligence
C Profesor José García Santesmases sn, 28040 Madrid, Spain
{angel.dba, angel.serrano, manuel.freire, imartinez, balta}@fdi.ucm.es

Abstract—The Learning Analytics (LA) discipline analyzes educational data obtained from student interaction with online resources. Most of the data is collected from Learning Management Systems deployed at established educational institutions. In addition, other learning platforms, most notably Massive Open Online Courses such as Udacity and Coursera or other educational initiatives such as Khan Academy, generate large amounts of data. However, there is no generally agreed-upon data model for student interactions. Thus, analysis tools must be tailored to each system's particular data structure, reducing their interoperability and increasing development costs. Some e-Learning standards designed for content interoperability include data models for gathering student performance information. In this paper, we describe how well-known LA tools collect data, which we link to how two e-Learning standards – IEEE Standard for Learning Technology and Experience API – define their data models. From this analysis, we identify the advantages of using these e-Learning standards from the point of view of Learning Analytics.

Keywords: Learning Analytics, e-Learning Standards, SCORM, Experience API, educational data mining

I. INTRODUCTION

Companies like Amazon, Facebook and Google study users' activity within their websites to adapt and improve the algorithms that sustain their business logic. Gathered data is used to adapt user interfaces, recommend new products or target advertising, among other tasks. Many other companies are using technologies such as web-analytics or business intelligence to better understand their customers and to improve their business. Data-driven approaches hold great promise towards improved decision-making. The e-Learning community is now beginning to apply these analysis techniques within a new trend called Learning Analytics (LA). This discipline gathers and analyzes educational data with different purposes such as seeking patterns in the learning process and trends or problems in student performance.

The educational experience is increasingly taking place within Learning Management Systems (LMS) deployed by educational institutions [1]. Within these computer-mediated environments, students interact with forums, on-line exercises, digital learning tools, games, and other types of digital content. Each individual interaction can provide one or more datapoints, and the system can collect a huge amount of data on student actions, courses and learning tools [2]. In addition to traditional LMS, Massive Open Online Courses (MOOC) such as Udacity, Coursera or educational systems such as Khan Academy, are increasing their acceptance as learning tools. These open courses also store large amounts of data about the students' performance. LA uses data mining and visual analytics techniques to derive actionable information from gathered data (both from LMS and MOOC). The goal is to detect and address learning problems, assess students, and predict learning results. Students can use these analysis results as guidance and self-awareness tools; teachers can use them to identify issues and try to tackle them; and schools can use results as a domain-specific variant of Business Intelligence [3], [4].

LA is a wide field that covers different aspects. Campbell deBlois and Oblinger identify five LA steps [5]: collect, report, predict, act and refine. Each step builds on the previous ones; therefore, data collection is critical to successful analysis. However, current LMSs lack standardized data structures; thus, LA tools tend to be tied to specific implementations of LMS and databases. This has a number of negative consequences: data gathered across different LMSs, or even different versions of the same LMS, are hard to move and compare; cross-institution data comparison is impeded, due to installation-specific data model differences; and LA tool adoption remains relatively low.

Many educational organizations and content-development enterprises have combined efforts to develop standards for e-Learning content interoperability [6]. Among these initiatives, some have addressed the problem of student performance data interoperability.

- The IEEE Standard for Learning Technology standard's family (from now on SLT family) provides a complex data model structure for tracking information on student interactions with learning content (IEEE 1484.11.1 [7]); additionally, an API allows digital educational content and the LMS to query and share collected information (IEEE 1484.11.2 [8]).
- The Experience API [9] is a very recent specification that presents a flexible data model. This specification adds the possibility of sharing tracking data among different LMSs.

The following organizations have partially supported this work: the Spanish Ministry of Science and Innovation (grant no. TIN2010 21735-C02-02); the European Commission, through the Lifelong Learning Programme (projects "SEGAN Network of Excellence in Serious Games" - 519332-LLP-1-2011-1-PT-KA3-KA3NW and "CHERMUG" - 519023-LLP-1-2011-1-UK-KA3-KA3MP) and the 7th Framework Programme (project "GALA - Network of Excellence in Serious Games" - FP7-ICT-2009-5-258169); the Complutense University of Madrid (research group number GR35/10-A-921340) and the Regional Government of Madrid (eMadrid Network S2009/TIC-1650).

Although the main focus of both standards is content interoperability, rather than improved data analysis, their adoption would have a profound (and highly positive) impact for LA tool users and designers.

In this paper, we present a study about how to use these standards for the "data collection" step in LA: extraction, storage in a concrete structure and sharing among different tools and systems.

This paper is structured as follows: in Section II, we identify current issues related to data collection and describe how several well-known LA tools try to overcome them; from this analysis, we propose a set of guidelines for a general LA data model. Section III analyzes the details of two of the best-known e-Learning standards for student data interoperability: the IEEE Standard for Learning Technology and the Experience API. In Section IV, we detail their implications from an LA point of view. Finally, Section V provides the conclusion and future work.

II. Current Issues Related to Data Access

There are several tools to analyze student experience. Most are internal to particular LMS, while a select few are external. However, there is a lack of integrated toolsets for comparing learner performance between different sources, or even individual student results with peer results. In order to improve the effectiveness of the toolsets, according to Siemens et al. [14], "analytics need to be broad-based, multi-sourced, contextual and integrated". To fulfill this vision, a large number of issues should be tackled, ranging from data access and acquisition to statistical modeling or network relationships.

LA and educational data mining are in their initial steps, and analysis tools are slowly appearing. Most of these tools are tightly coupled to specific systems, such as LMSs and MOOCs, because they rely on direct access to the educational systems' internal data-structures to perform analyses. Prior to any analysis, data must be collected, cleaned, and normalized to fit the LA's expected structures, in a process known as Extract, Transform and Load (ETL). To enable LA over heterogeneous, distributed environments, ETL tools should be able to collect data from different data repositories, via API calls or RSS feeds.

The following list describes the type of data gathered and analyzed by some well-known LA tools.

- SNAPP [10] uses social network analysis over discussion forum posts to categorize students into several profiles, such as engaged students, disconnected (or "at risk") students, or information brokers. SNAPP relies on forum interactions: posts written, messages replied to, and topics opened.
- LOCO-Analyst [11] provides teachers with feedback of web-based courses. It identifies the most relevant parts of courses and provides course content statistics (content-derived as tag-clouds). It can also relate course topics mentioned in forums with course parts. The tool relies on user access to the different course resources, and the time spent on each of them. It also reads forum contents and course materials.
- Course Signals [12] is integrated into the Blackboard LMS. It analyzes individual student performance to predict which students are at risk of performing poorly. Risk status has four components: performance, effort, prior academic history, and student characteristics. Two types of data are considered: static, such as prior history (e.g., academic preparation, high school GPA, standardized test scores) or student characteristics (age, residency or credits attempted); and dynamic data, such as course grades to date, or amount of interactions with the LMS as compared to peers.
- The Desire2Learn LMS includes an LA tool named Student Success System. This tool is also focused on high-risk student detection, to enable early intervention. The system relies on raw data for its analysis, including student grades, login frequency, discussion posts, and results and number of attempts in quizzes.
- Pittsburgh Science of Learning Center (PSLC) DataShop [13] is a repository of data extracted from different learning courses, most notably MOOCs offered by the Open Learning Initiative [14]. The PSLC DataShop's goal is to help in the development of standards for anonymized student data interoperability and interchange. According to PSLC DataShop, MOOCs split student progress into Knowledge Components, in which students can either succeed or fail. Every lesson contains several knowledge components, and the MOOC records the outcomes for each of them.
- Khan Academy [15] is a non-profit educational organization that supplies free web-based micro lectures via online video tutorials. Khan Academy records performance in all of the course exercises attempted by the students. Students can see their overall results, and teachers have an overview of the students' progress and of the exercises with weaker performance.

Most of the current LA tools use, as basis for their analysis, concrete actions performed by the students in the educational environments or learning tools, with or without additional variables associated with those actions that give a more detailed insight. Some tools also integrate student characteristics into their analyses, such as academic information and personal data (Table I). It is worth noting that LMS and MOOCs platforms rely on additional web-analytics data (page views, times, etc.) for their analysis.

The data tokens used by the previous tools can be classified into two categories:

- *Student-performed actions with a given outcome*: For example, a student viewed a resource during a certain time; finished an activity with a given result; or completed a quiz with a percentage of correct answers.
- *Student profile*: age, interests, gender, residency, etc.

With these two basic categories, and using suitable means of aggregation and summarization, we can build more complex data. For example, from a "student wrote a post" token, we can derive the total number of posts written in a forum by this student, by a group of students, or in a course. From an LA

perspective, it could also be interesting to know about the history of the student, that is, prior academic results. This information can be accessed by gathering data of the first category.

TABLE I. ANALYTICS TOOLS AND THEIR DATA FOR ANALYSIS

Analytic Tool	Platform	Data for analysis
SNAPP	External tool	Forum activity
LOCO-Analyst	External tool	Resource views, resource contents, forum contents
Course Signals	LMS	Student age, residency, credits attempted, academic history, course grades to date, interactions with the LMS
Desire2Learn Students Success System	LMS	Student grades, login frequency, discussion posts, results and number of quiz attempts.
Open Learning Initiative	MOOCs	*Knowledge Components* achieved and failed
Khan Academic	MOOCs	Performance in exercises

III. E-LEARNING STANDARDS FOR TRACKING DATA INTEROPERABILITY

The interoperability of educational content in different systems was driven by the expansion of Technology Enhanced Learning (TEL) in both industry and educational institutions. The high costs of developing educational content that would later be tied to particular software (and even hardware), hindering distribution and reuse, prompted several organizations to look for alternatives. After analyzing patterns and guidelines to unify content distribution, these organizations created e-Learning standards that enabled content interoperability.

Student performance data interoperability and services to share these data are among the multiple features provided by these standards. This is the case of the IEEE SLT family and the new Experience API specification. In this section, we analyze the potential of each initiative in terms of the kind of data that it stores and the communication services it provides.

A. IEEE Standard for Learning Technology

The Aviation Industry Computer-Based Training Committee (AICC) developed many techniques for both hardware and software standardization of TEL. The most important among them by their impact and acceptance is the Content Management Instruction (CMI) specification, a set of guidelines for interoperability between web courses and the LMS. The CMI provides both a data structure for student interactions with learning contents as well as an API for managing these data. This work was the basis for the IEEE Learning Technology Standards Committee when developing the data exchange model (IEEE 1484.11.1) and communication specification (IEEE 1484.11.2). Both standards are part of the IEEE Standards for Learning Technology (IEEE 1484.11) and are included in the widespread Sharable Content Object Reference Model (SCORM) specification.

The data exchange model provides a large range of fields for storing different aspects of student interactions. First, there is a set of fields intended for the storage of general information on a student's degree of progress in a given activity. These fields are "End State" (`cmi.completion_status`) and "State of Success" (`cmi.success_status`). In addition, the data model can store an overall student performance (`cmi.score.raw`) on a range of values (`cmi.score.min` and `cmi.score.max`).

The "Objectives" field (`cmi.objetives`) links the completion of different parts of the educational content to specific learning objectives. Data stored includes degree of completion and success (local to the target), progress measurement, score, and how much each objective counts towards the final grade.

The IEEE data exchange model defines a field consisting of a list of records named "Interactions" (`cmi.interactions`) to store fine-grained information regarding student interaction with the learning content. This field can store, for example, a student's answers to a set of questions, or her specific actions within a task. The field is a set of records: each record includes not only the student answer and the result (i.e., whether or not the student was right), but also the type of interaction (e.g., true-false, relationships between elements of two groups), the correct answer as set by the instructor, or the weight of each interaction to the final grade. For greater expressiveness, multiple correct answers can be included, each with a specific degree of correctness. Additionally, the "interactions" field also supports tagging particular entries with identifiers that link them to sets of related learning objectives.

An important feature of "Interactions" is the possibility of storing data in two different ways: as a journal, by adding a new record to the set, or by status, storing only one copy of each interaction. The first mode allows for detailed storage of actions performed by the student while the second allows the final state of each interaction to be stored (as each interaction would overwrite the previous state).

Finally, the IEEE data exchange model includes fields that allow the storage of information related to the state of the educational content; this allows students to resume activities at the point in which they were interrupted. The data model also includes a (`cmi.comments_from_learner`) field to collect student feedback on their educational experience.

The IEEE learning content interoperability defines an API for sending and requesting data between the educational content and the LMS. This API contains methods to initialize and finalize the communication, and to store and retrieve data. This API is only accessible by the educational content. From the point of view of LA, the fact that other systems cannot use it (such as LMS, content repositories, reporting tools, or LA tools) is a significant drawback.

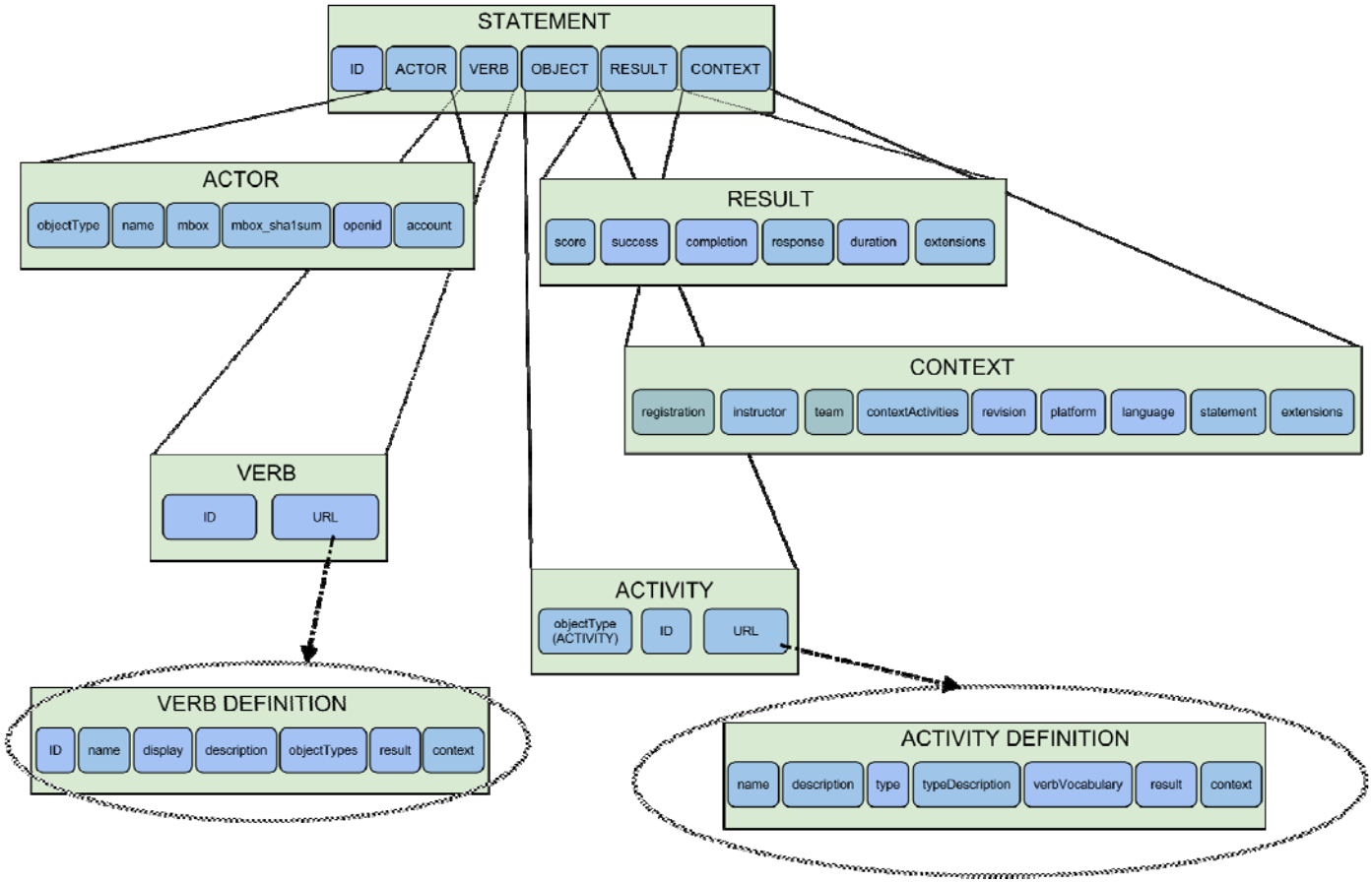

Figure 1. An overview of the Experience API data model

B. Experience API

The Experience API (formerly known as Tin Can API) is a new e-Learning specification under development by ADL and Rustici Software, and it is considered the "new generation of SCORM". ADL has released draft versions for early adopters (i.e., the last specification is v0.95) to elicit feedback from users before releasing the final version. The Experience API is focused on defining an interoperable data model for storing data about students' learning experience and an API for sharing these data among systems. It also addresses some of SCORM's shortcomings regarding data access.

The central element in Experience API is the Learning Record Store (LRS). The LRS can reside inside the LMS or in an independent server. A specific module for data storage allows learning tools to be decoupled from the LMS, and to send information whenever they have connectivity (rendering permanent connectivity no longer necessary). This allows episodically-connected learning activities, such as those delivered on mobile devices, collaborative tools, virtual worlds, or simulations, to report information on the learning experience. At any time (including after the end of the experience), activities can send in their collected data over the Experience API web service. The service is available regardless of whether activities are taking place inside or outside the LMS. In addition to providing improved data collection, the Experience API also allows different LMS, servers, web applications or reporting tools to share tracking information.

The Experience API data model takes, as a starting point, the concept of Activity Streams [16], where the users' activity is stored as statements: "I did this". The Experience API data model extends this idea to track all aspects of the learning experience. Thus, Experience API statements have the following structure:

<actor> <verb> <object>, with <result>, in <context>

The `actor` (usually a learner), `verb` and `object` elements are mandatory and can be complemented with `result` and `context` elements. The most important feature of the Experience API data model is the flexibility provided not only by the statement's structure, but also by its elements.

Students can interact with educational content via different systems or tools. For this reason, the `actor` element allows different IDs of the same student to be used for each system, instead of keeping a centralized registry of unique users for LRS purposes. In addition, since actors can be represented using other systems' IDs, a measure for anonymity is provided – true user identities are only available at the system where each ID is maintained.

The `verb` element is a key part of statements, because it describes the action performed by the student. A `verb` is not a

simple string, as it also includes a URL where this `verb` is defined. A definition includes name, description, and the recommended best practices for its usage. Although some verbs can be used for many different learning activities (e.g., "experienced", "attempted", "failed", "passed"), specific communities of practice (such as the serious games community) can extend the list of verbs or clarify the verb's meaning.

The `object` element represents "who" or "what" experienced the action defined in the `verb`, and can therefore be an `actor` or a learning activity (for instance, in the record "Student X experienced webinar Y"). Learning activities also include a URL pointing to their definition, which can include other information such as a description of the learning activity, verbs that can be used with the activity, the fields of applicable `result` elements that can be filled or best practices cases.

The `result` element provides an outcome to the statement. It includes score, success and completion fields. The `context` field adds extra information to the statement. For example, `context` could include information about the relationship of the activity with other activities, its position in a learning sequence, or the name of the instructor, among others. In order to improve the flexibility of the data model and allow for the extension of some elements to additional scenarios, activity definitions, `context` and `result` elements have an extension field which can contain any pair of key/value data.

The Experience API also includes a set of REST services for data transfer (including POST, PUT, GET and DELETE). The services do not only allow sending statements to the LRS, but also information about activities and actors. The Experience API uses either OAuth or HTTP Basic Authentication to authenticate access to LRS services. Therefore, the LRS API can be accessed by any system or digital content with the necessary credentials.

IV. Using e-Learning Standards for Learning Analytics Purposes

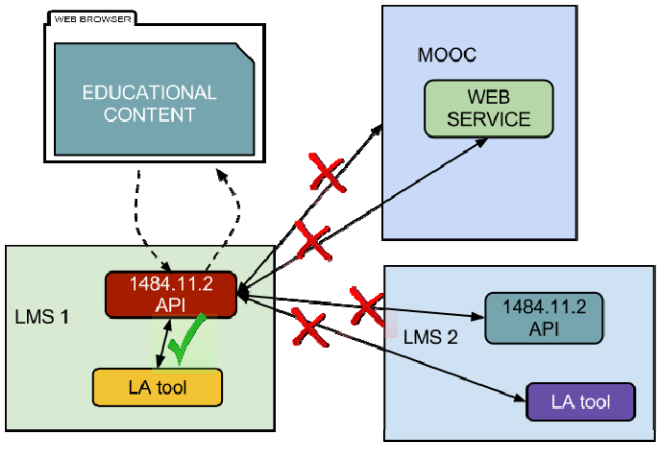

Figure 2. IEEE Standard for Learning Technology family: LA tools can only benefit from data stored in their host LMS; red crosses indicate communications that are not covered by the standard, and must be implemented *ad-hoc* for each LA. In addition, the educational content is tied to the LMS.

The main goal of e-Learning standards is the interoperability of digital contents and learning tools among different e-Learning systems (e.g. LMS and MOOCS). When these standards describe models for student interaction data, LA stands to benefit greatly from their adoption. On the one hand, development costs are greatly reduced, and investments are made future-proof, since tools will continue to work as long as they adhere to the standards. On the other hand, the decoupling of LA tools from specific systems facilitates data reuse and broadens the pool of data that can be analyzed and explored. In addition, stable data sources and structures enabled by standardization would allow LA tool developers and researchers to focus on other open issues, such as better statistical analysis and visualization.

To maximize the benefits of LA, e-Learning standards must first meet certain requirements. First, the data model structure should be able to represent the two categories identified in Section II: actions performed by students with the associated outcomes, and student profiles. Second, the data structures should provide API methods to access and share data among systems, data repositories and reporting tools.

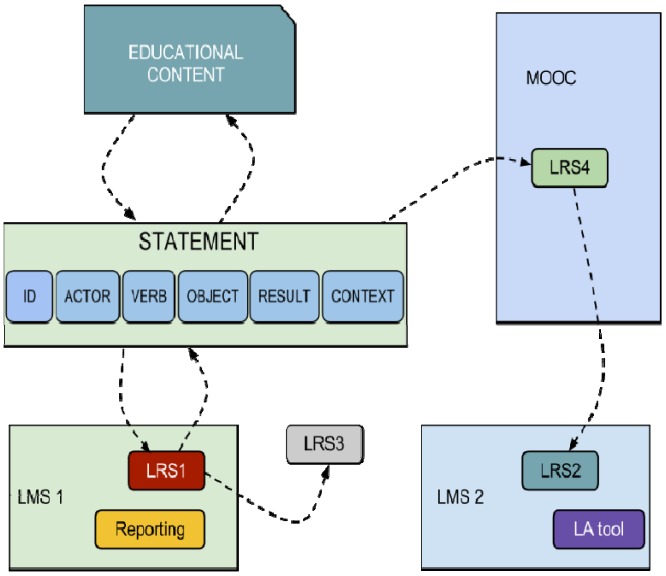

Figure 3. Experience API: statement structure and data flow. The statements can be sent to different LRS at the same time and can be shared among LRS. In addition, the educational content is not tied to particular LMS.

The IEEE data exchange model proposes a set of fields to store data about learner performance. Although the fields in this data model can certainly store a wealth of useful information for LA tools, there are still open issues. First, this data model does not explicitly use "verbs" (well-defined, standardized actions). Building appropriate verbs would be easy for certain fields; other fields could be interpreted with different verb meanings. This ambiguity hinders universal usage of stored data and could lead to inaccurate analysis. Moreover, certain fields in the data-model mix information on student actions and outcomes, leading to an additional source of ambiguity. Second, certain learning experiences cannot be represented within the current model. This is partially addressed

with a data model extension for the IEEE data exchange model: IMS Shareable State Persistence (IMS-SSP) [17]. However, this extension can also to lead to misinterpretations, due to the verb-ambiguity problem. Finally, the IEEE communication specification does not allow access to data from other systems or tools (Figure 2).

The Experience API provides a better fit for the requirements of LA tools. On the one hand, it presents a flexible data model that allows for representing student actions in a univocal way (roughly exemplified as *"student X performed action Y with outcome Z [in context W]"*), with optional context information. Furthermore, student privacy and anonymity can be preserved, since data records do not require unique or identifiable user IDs. On the other hand, the Experience API runtime communication provides access to the data from other systems. The decoupled nature of the LRS allows LMSs reporting tools, LA tools, and any other system with appropriate credentials to store or access tracking data. However, the Experience API lacks specific support for any student profile information.

V. Conclusion and Future Work

The growth of the field of Learning Analytics is a direct consequence of the increased use of e-Learning systems, in an attempt to harness the large amounts of data that these systems generate with educational or administrative purposes. Among the remaining research issues, there is a need for common structures into which these data can be stored and associated services to query it. After analyzing how a set of well-known LA tools extract data from different systems (e.g. LMSs and MOOCs), we identified two common structures: actor-action-object statements for dynamic data and static student profiles. In this paper, we have analyzed how current e-Learning standards for interoperability of student performance data, more specifically the IEEE Standards for Learning Technology and Experience API, can help in the development of LA tools.

On the one hand, the IEEE Standard for Learning Technology, included in the SCORM specification, has been widely adopted. However, actions in the data model are potentially ambiguous, and the API runtime services do not allow external systems to retrieve the data. The statement structure of the Experience API data model and the decoupled nature of the LRS make it a better choice for accessing and sharing data. Nonetheless, the lack of specific support for student profiles is hindering its adoption as a universal solution. However, this can also be seen as a calculated tradeoff; between increased data-sharing (without personal profiles) and increased expressive power (at the expense of potential privacy breaches), its designers appear to have chosen the safe route. Note that Experience API is currently under development and is subject to change. The current analysis is based on the available published draft.

As future work, we are currently implementing Experience API-compliant data access for our own LA tools [18]. We plan to use it to collect and analyze data from educational games.

Acknowledgment

Thanks to Kathy Come for her detailed English editing and review.

References

[1] M. F. Paulsen, "Experiences with Learning Management Systems in 113 European Institutions," *Educational Technology & Society*, vol. 6, no. 4, pp. 134–148, 2003.

[2] EDUCASE, "7 Things You Should Know About Analytics," 2010.

[3] R. Kop, "The Challenges to Connectivist Learning on Open Online Networks : Learning Experiences during a Massive Open Online Course," in *European Distance and E-learning Network annual Conference 2010*, 2010, p. Paper H4 32.

[4] M. De Laat, "Networked Learning," Utrecht Universiteit, 2006.

[5] J. P. Cambpell, P. B. DeBlois, and D. G. Oblinger, "Academic Analytics A New Tool for a New Era," *EDUCAUSE Review*, vol. 42(4), no. August 2007, pp. 42–57, 2007.

[6] N. Friesen, "Interoperability and Learning Objects : An Overview of E-Learning Standardization," *Interdisciplinary Journal of Knowledge and Learning Objects*, vol. 1, pp. 23–31, 2005.

[7] IEEE, "IEEE 1484.11.1, Draft 5 Draft Standard for Learning Technology— Data Model for Content Object Communication." 2004.

[8] IEEE, "IEEE 1484.11.2/D2 Draft Standard for Learning Technology—ECMAScript Application Programming Interface for Content to Runtime Services Communication." 2003.

[9] ADL-Co-Laboratories, "Experience API Version 0.95."

[10] S. Dawson, A. Bakharia, and E. Heathcote, "SNAPP : Realising the affordances of real-time SNA within networked learning environments," in *Learning*, 2010, pp. 125–133.

[11] L. Ali, M. Hatala, D. Gašević, and J. Jovanović, "A qualitative evaluation of evolution of a learning analytics tool," *Computers & Education*, vol. 58, no. 1, pp. 470–489, 2012.

[12] M. D. Arnold, K. E. & Pistilli, "Course Signals at Purdue: Using learning analytics to increase student success.," in *Proceedings of the 2nd International Conference on Learning Analytics & Knowledge. New York: ACM.*, 2012.

[13] J. Koedinger, K., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., Stamper, *A data repository for the EDM community: the PSLC datashop*. CRC Press, 2010.

[14] F. Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. and R. . R., Duval, E., Verbert, K., Baker, ": Open learning analytics: an integrated and modularized platform: Proposal to design, implement and evaluate an open plat- form to integrate heterogeneous learning analytics techniques," 2011.

[15] S. Khan, "Khan Academy," *Educational Technology*, 2011. [Online]. Available: http://www.khanacademy.org/.

[16] ActivityStreamsWorkingGroup, "JSON Activity Streams 1.0," 2006.

[17] IMS Global Consortium, "IMS Shareable State Persistence, Version 1.0 Final Specification," 2004.

[18] A. Serrano, E. J. Marchiori, A. del Blanco, J. Torrente, and B. Fernandez-Manjon, "A framework to improve evaluation in educational games," in *Proceedings of the 2012 IEEE Global Engineering Education Conference (EDUCON)*, 2012, pp. 1–8.