Data Models in Learning Analytics

Vlatko Lukarov, Dr. Mohamed Amine Chatti, Hendrik Thüs, Fatemeh Salehian Kia, Arham Muslim, Christoph Greven, Ulrik Schroeder

Lehr- und Forschungsgebiet Informatik 9
RWTH Aachen University
Ahornstrasse 55
52074 Aachen
lukarov@cil.rwth-aachen.de
chatti@informatik.rwth-aachen.de
thues@cs.rwth-aachen.de
fatemeh.salehian@rwth-aachen.de
muslim@cil.rwth-aachen.de
greven@informatik.rwth-aachen.de
schroeder@informatik.rwth-aachen.de

Abstract: There are many different ways and models how to characterize usage data to enable representation of user actions across learning management system, and systems in general. Based on this data, learning analytics can perform different analysis and provide personalized and meaningful information to improve the learning and teaching processes. There is a variety of usage data formats that are already successfully used in exiting systems. These different usage data formats have their advantages and disadvantage that have to be considered when using them in the context of learning analytics. In this paper, several usage data formats are presented and analyzed in the context of learning analytics to help in choosing the best suiting usage data model.

Keywords: usage data models, learning analytics

1 Introduction

Learning Analytics as young and emerging field has many definitions. If one takes a closer look at these definitions, she will notice the definitions have differences in the details. One will also notice that these definitions share an emphasis on converting educational data into useful actions to foster learning. Additionally, it is noticeable that these definitions do not limit Learning Analytics to automatically conducted data analysis. Learning Analytics is so far, data-driven approach, and as such uses various sources of educational data. These data can come from (but not limited to): centralized educational systems, distributed learning environments, open data sets, personal learning environments, adaptive systems/ITS, web-based courses, social media, student information systems, and mobile devices. These data sources in the background have

centralized educational systems. These are in essence learning management systems, such as Blackboard, Moodle, L^2P, or Ilias. These learning management systems accumulate large logs of students' activities and interaction data. Additionally, these systems are often used in formal learning settings to enhance traditional face-to-face teaching methods, or to support distant learning. The user generated content, facilitated with ubiquitous technologies, has led to vast amounts of produced data by students across learning environments, and systems [CDST12].

In short, the learning data can and should come from formal and informal channels, because learning and knowledge creation is often distributed across multiple media and sites in networked environments [SR11]. The challenge is how to aggregate and integrate raw data from multiple and heterogeneous sources, often available in different formats, to create a useful educational data set that reflects the activities of the learner, hence leading to better Learning Analytics results.

2 Data Models

The user activities and their usage of data objects in different applications is called Usage Metadata. Today, there is a growing number of data representation formats for usage data. These are not just simple logging files, but they focus on the users' activities. This paper first presents the four most commonly used data representations, namely Contextualized Attention Metadata, Activity Streams, Learning Registry Paradata and NSDL. Then it is intended to provide IMS specifications of how learning systems should capture and share data around learning interactions. This paper concludes by suggesting for improvement of the learning context data model.

2.1 Contextualized Attention Metadata (CAM)

Contextualized Attention Metadata (CAM) allows monitoring user interactions with learning environments. The focus has moved from the user and the data object to the event itself. This means that events can have flexible set of attributes.

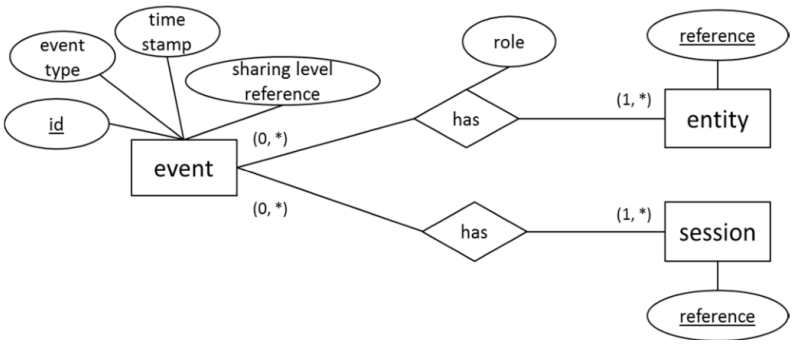

Figure 1: CAM Scheme

Figure 1 depicts the latest version of CAM scheme. This scheme stores basic information about an event. Other information for each *event* is stored as *entities*. Due to the simple and abstract scheme, a lot of information has been removed to the *role* attribute. This needs to be defined from the starting point. For instance, some sample values of *role* can be sender, receiver, context, writer, forum, thread. It also requires rules to be enforced on the instances of *role* attribute, that is, if the role attribute is "forum", there needs to be exactly one related *entity* with the role attribute "*writer*" and at least one with the value "*message*". *Session* defines time span in which the event occurred. This scheme with a simple and flexible representation can be suited for different learning platforms, but it requires defining rules and constraints to make the model more clear and consistent. The information can be stored in different formats such as JSON, XML, RDF, or in relational database [NSW12].

2.2 Activity Streams

An Activity Stream (Figure 2) is a collection of one or more individual activities carried out by users. Each activity comprises of certain attributes. Figure 2 shows the activity streams scheme. An *activity* has three properties e.g. *actor*, *object*, and *target*. Each property is an *object* in activity stream format. The *verb* attribute plays the same role as *event type* in the CAM scheme. It describes an action which is done in the learning activity. Additionally, every object that is within an Activity Streams object can be extended with properties not defined by the core definition and specification and this way a lot of flexibility is provided [NSW12].

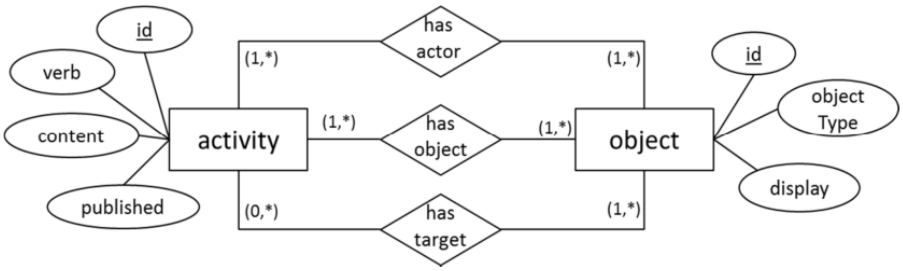

Figure 2: Activity Streams Scheme

2.3 Learning Registry Paradata

Learning Registry Paradata (Figure 3) is an extended version of Activity Streams for storing aggregated usage information about resources. The three main elements of Learning Registry Paradata are *actor*, *verb*, and *object*. The verb refers to a learning action and detailed information can be stored [NSW12].

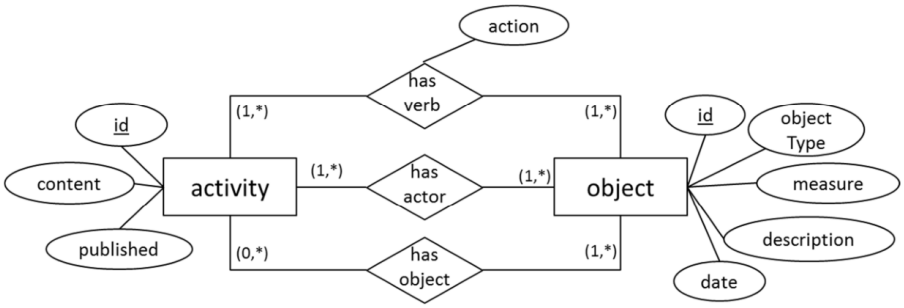

Figure 3: Learning Registry Paradata scheme

2.4 NSDL Paradata

This data format (Figure 4) collects aggregated data about resources such as downloaded or rated resources. Despite the fact that other usage data formats are event centric this format is object-centric. The main element is the *usageDataSummary* which comprises all available usage statistics/information about a resource using five different types of values e.g. *integer/float*, *string*, *rating type*, *vote type*, *rank type*.

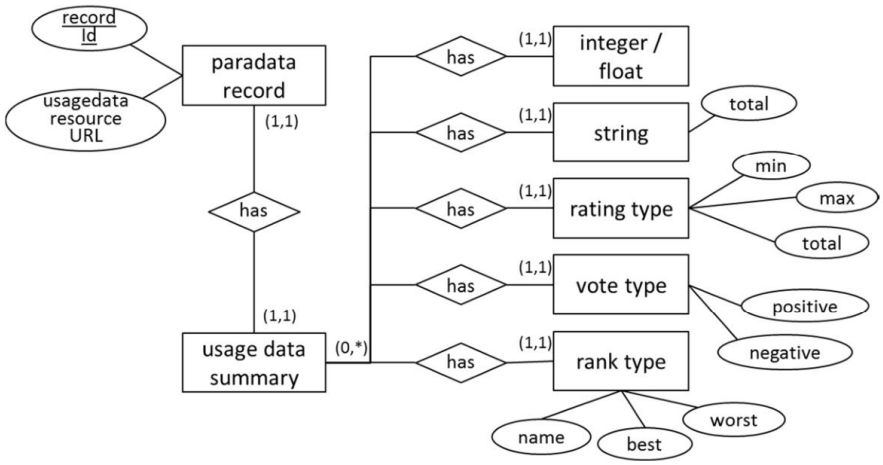

Figure 4: NSDL Paradata scheme

Integer/float shows the number in which certain action is performed on the resource e.g. "downloaded" or "rated". *String* can be a textual value such as comment. A *rating type* represents an average rating value in respect to certain criteria, for instance, usability of

the resource. The *vote type* and *rank type* represents the interest rate on a specific resource. It is worth noting that the extensive version of NSDL Paradata contains more details regarding *usageDataSummary* such as audience of used resource, and the subject of the resource [NSW12].

2.5 IMS Specifications of Learning Measurement for Analytics (IMS Caliper)

IMS defines a learning measurement framework, Caliper. IMS Caliper is built around these three concepts: IMS Learning Metric Profiles, IMS Learning Sensor API, and Learning Events IMS LTITM/LIS/QTITM leverage and extensions. The idea behind learning metric profile is to define the structured collection of learning activity metrics which represents measurements specific to actions within each genre of activity. Most learning activities can be grouped into one or more classes e.g. reading, assessment, media etc. In addition, there are Foundational Metrics such as engagement, and performance. Figure 5 depicts a sample of IMS Caliper scheme connected with different IMS Metric Profiles [IMS13].

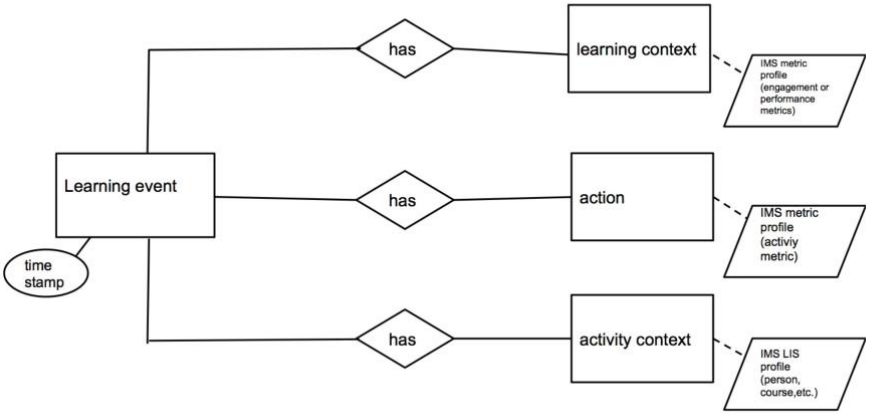

Figure 5: LMS Caliper scheme

2.6 Learning Context Data Model

The new L²P follows a student centred approach and focuses on customizability, extensibility and mobility. So, there exists various delivery learning environments, and the data model has to be defined in order to collect all the required information as well as to be independent of each learning platform. The learning context data model is based on CAM representation. To answer the question of which abstraction level is suitable for this data model requires considering two points. First, we have to take into account which type of learning activities should be filtered. Second, we should consider how to maintain the semantic of context information while they are coming from different platforms such as mobile or web based. The proposed data model is shown in Figure 6.

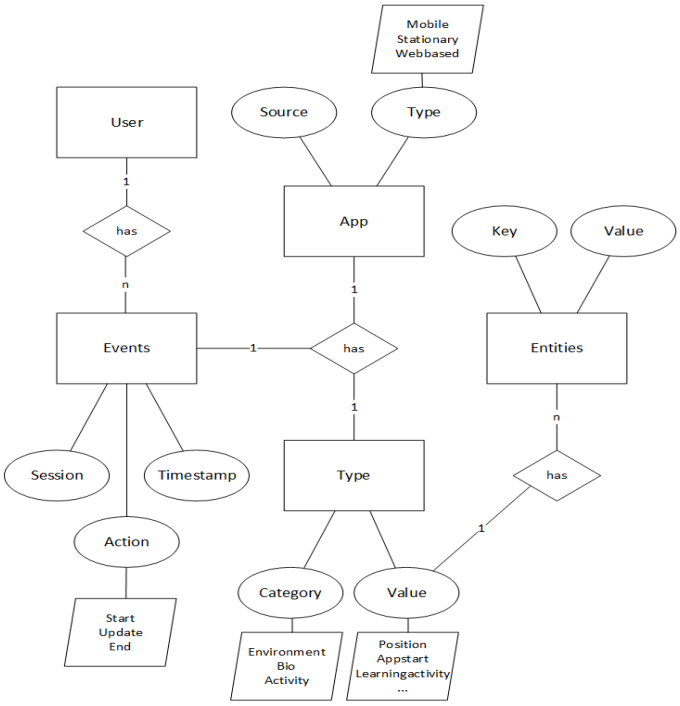

Figure 6: Learning context data model scheme

3 Comparison

In section 2 we have presented 6 different data models. As it can be seen in Table 1 they are divided into four main categories, depending on which element is the main element in the data model. The CAM data model and the IMS Caliper data model focus their model on the learning event. For the Activity Streams data model and the Learning Registry Paradata the main element is the learning activity. This is one level of abstraction more detailed from the event of the CAM, or IMS Caliper. NSDL Paradata focuses on the object that presents the summary of the usage data. The last one is centered on two elements which are both the user and the event. We think that the event is important, but also it is the user who triggers the events, and this information is crucial, in order to keep the semantic knowledge from where the user is accessing the learning system (mobile or desktop). Based on this, we can better personalize and better amend the analytics results to help both teachers and students. Another point to consider is the level of abstraction. While CAM and Activity Streams (and their variations) are very abstract, the IMS Caliper with the IMS Metric Profiles is very detailed and complex especially when it comes to single users. There should be a balance between the level of abstraction and the complexity of the data models. The data models for activity aggregation might not be suitable for personalized results concerning Learning Analytics.

Event Centric	
Contextualized Attention Metadata (CAM)	
Main element	Event
Other elements	Entity, Session
IMS Caliper	
Main element	Learning event
Other elements	Activity Context, Action, Learning Context
Activity Centric	
Activity streams	
Main element	Activity
Other elements	Actor, Target, Object
Learning Registry Paradata	
Main element	Activity
Other elements	Actor, Verb, Object
Object Centric	
NSDL Paradata	
Main element	usageDataSummary
Other elements	Integer/float, string, rating type, vote type, rank type, paradata record
User Centric	
Learning Context Data Model	
Main element	User, Event
Other elements	App, Type, Entities

Table 1: Data Models Comparison

4 Conclusion

We reviewed six prevalent data models which can be used to represent usage data for learning analytics. We have provided schemas, and described their properties. These data models have been created with purpose to serve analytics (recommender systems, data mining, learning analytics). Researchers, developers, system designers must know their strengths, and their weaknesses when using them to manipulate and represent usage data in their respective applications. As mentioned in the review, one should distinguish what is the purpose of his learning analytics tool, and accordingly choose the data model. As balanced model that is abstract enough, but also provides enough detailed information could be taken the learning context data model. However, one should not take these data models for granted and complete, but rather work on additional elements that will better organize the data, thus making the analytics results more precise.

References

[CDST12] Chatti, M.A.; Dyckhoff, A.L.; Schroeder, U.; Thüs, H.: A Reference Model for Learning Analytics. In International Journal of Technology Enhanced Learning 2012 Vol. 4 No. 5/6

[NSW12] Niemann, K.; Scheffel, M.; Wolpers, M.: An Overview of Usage Data Formats for Recommendations in TEL. In Proceedings of the 2nd Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2012)

[SR11] Suthers, D.; Rosen, D.: A unified framework for multi-level analysis of distributed learning. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge. NY, USA: ACM New York. (pp. 64-74).

[IMS13] IMS Global Learning Consortium Inc.: Learning Measurement for Analytics Whitepaper 2013 http://www.imsglobal.org/IMSLearningAnalyticsWP.pdf

[TCYPKMS12] Thüs, H; Chatti, M.A.; Yalcin, E; Pallasch, C; Kyryliuk, B; Mageramov, T; Schroeder, U: Mobile Learning in Context. In International Journal of Technology Enhanced Learning 2012 Vol. 4 No. 5/6