An Overview of Usage Data Formats for Recommendations in TEL

Katja Niemann, Maren Scheffel, Martin Wolpers

Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

{katja.niemann, maren.scheffel,
martin.wolpers}@fit.fraunhofer.de

Abstract. Recently, a number of usage data representations have emerged that enable the representation of user activities across system and application boundaries. Based on these user activity data, systems can adapt to the users and provide personalized information. A lot of usage data representation formats are already successfully used in real world applications. However, dependent on the purpose, the formats show different advantages and disadvantages one must consider when choosing a format for a system. In this paper, we will present the four most commonly used data representations, namely Contextualized Attention Metadata, Activity Streams, Learning Registry Paradata and NSDL to alleviate the selection of a suitable format.

Keywords: usage data formats, technology enhanced learning

1 Introduction

Attention or Usage Metadata represent the activities of users and their usage of data objects in specific applications. Aggregating and analysing the usage data provides the basis for advanced user support systems, e.g. learning recommendation or self-reflection support. Furthermore, usage data can be employed for annotating data objects with information about their users and usages, thereby rendering possible object classifications according to use frequency, use contexts and user groups [1], [2] [3].

Particularly in the domain of learning analytics (see [4], [5] and [6] for more information on learning analytics) and educational data mining, usage data provide the basis for learning support systems. For example, based on an analysis of usage data, irregularities of learning behaviour of students can be identified [7] and the results of corrective activities by the teacher can be monitored. Another example of the successful application of analysing usage data in learning settings is the reflection and comparison of learning activities among students of a learning group. Here, by playing back their learning activities, students compare themselves with their fellow students and identify how to improve their learning activities. A further example of the successful use of usage data are personalized recommender systems, e.g. in the domain of learning (see [8] for more details on recent learning recommendation systems).

Recently, a number of data representation formats for usage data have emerged. In contrast to simple logging files, these representations focus on the activities of users and not on those of a system. In this paper, we will present the most prominent examples, namely Contextualized Attention Metadata, Activity Streams, Learning Registry Paradata and NSDL Paradata.

2 Usage Data Formats

2.1 Contextualized Attention Metadata

The CAM scheme [9] was defined as an extension of Attention.XML [10] which is an early approach to capturing and storing attention metadata for single users. In the current CAM version[1], the focus has moved from the user and the data object to the event itself. This is due to the insight that not every event has a fixed set of attributes.

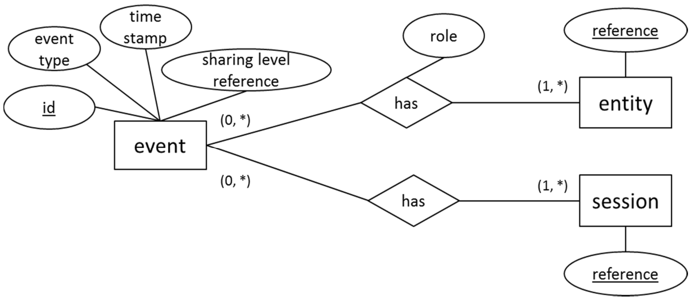

Fig. 1. CAM scheme

Additionally, only the basic information about an event is stored, e.g. the event type and the time stamp. All other information, e.g. metadata describing users or documents involved in the event, are linked. In this way, each entity and also each session can be described in a different and suitable way and no information is duplicated.

Fig. 1 shows the complete CAM scheme. The main element of a CAM instance is the *event* entry which comprises its *id*, the *event type*, the *timestamp*, and a *sharing level reference*. Examples for *event types* are "send", "update" or "select". The *sharing level reference* points to a description of the specific sharing level which describes the privacy related issues of the event. Depending on the event, various entities with different roles can be involved, e.g. when sending an e-mail, there is a person with the role sender, at least one person with the role receiver and a document with the role e-mail. Each event can be conducted in a session. A *session* can, for example, be the time between booting and shutting down a computer or the time between the login and logout of a user in a portal.

The current CAM scheme does not have fixed bindings so far. The information can be stored in XML, RDF, JSON or in a relational database, depending on the purpose of the data collection.

[1] https://sites.google.com/site/camschema/

2.2 Activity Streams

The Activity Streams specification [11] defines a format for single activities carried out by users. An Activity Stream is a collection of one or more individual activities. Usually, activities are serialized using JSON.

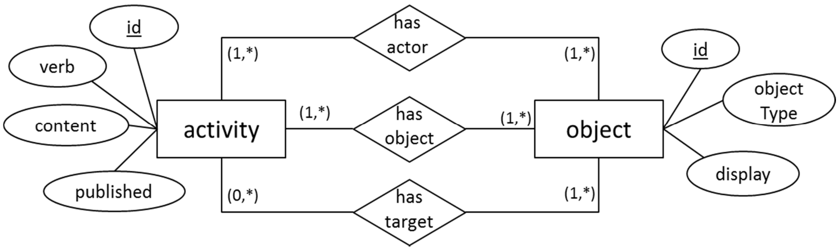

Fig. 2. Simplified excerpt of the Activity Streams scheme

Fig. 2 shows the core elements of the Activity Streams scheme. A single activity must at least contain a description of the entity that performed the activity (*actor property*) and the date and time at which the activity was published (*published property*). The Activity Stream Working Group recommends that an activity also contains a *verb*, an *object*, and an *id property*. The *verb* identifies the *action* that the *activity* describes (e.g. “accept”, “add”, “dislike” etc.), the *object property* describes the primary *object* of the *activity* (e.g. the watched movie or the sent e-mail) and the *id property* provides a unique identifier for the activity in the form of an absolute IRI (Internationalized Resource Identifier). The *target property* is optional and can be used if indicated by the verb. For instance, in the activity, "John sent an e-mail to Bill", “Bill” is the target of the activity.

The value of the *actor, object, and target property* respectively is an Activity Stream Object. An Activity Stream Object comprises several properties describing the object and should at least contain an IRI (*id property*) and a plain-text name for the object (*display property*). Additionally, it can contain others such as an *object type*. The Activity Base Schema [12] already defines *object types* to be used with Activity Streams, e.g. “alert”, “application”, “article”, etc. The *object types* are further grouped in six classes, i.e. audio and video objects, binary objects, events, issues, places, tasks. Depending on the class, *objects* may contain further properties, e.g. *startTime* and *attendedBy* for Events. Furthermore, any object within an Activity Streams object can be extended with properties not defined by the core Activity Streams’ specification to provide as much flexibility as possible.

2.3 Learning Registry Paradata

The Learning Registry Paradata format [13] is basically an extended and altered version of the Activity Streams JSON format. It was defined to store aggregated usage information about resources. The Learning Registry Paradata specification states explicitly that the Activity Streams format should be used if mainly individual actions are stored.

As for the Activity Streams, a basic LR Paradata statement consists of three key elements: *actor*, *verb*, and *object* (see Fig.3). The *actor* refers to the person or group that does something and is represented by a string or LR Paradata object (as defined later). The *verb* refers to the action that is taken. In its simplest form, it just contains the *action* name (e.g. "taught" or "viewed"), but it can also be specified in more detail, which is the main difference of the LR Paradata and the AS scheme. The *object* refers to the thing being acted upon using a string or a LR Paradata object.

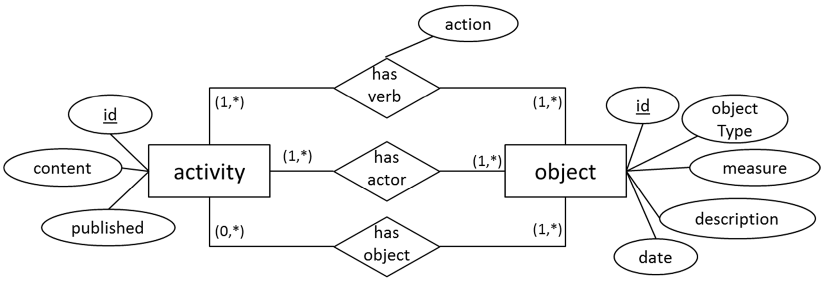

Fig. 3. Simplified excerpt of the Learning Registry Paradata scheme

A LR Paradata object may contain an *id*, an *objectType*, a *description* of the object, i.e. an array of *keywords*, a *measure* related to the object, a *date* and a *context*. Apart from *description* and *date*, each element can be represented by a string or by a JSON object without pre-defined scheme. The values of the elements depend on the *objectType*, which can be e.g. a person, a group, a learning resource, a LMS, etc. Within the *verb*, an action is specified that holds the verb's value (*action*), additionally, it can contain any element specified for the LR Paradata object [14], [15].

2.4 NSDL Paradata

The NSDL Paradata format was defined to capture aggregated usage data about a resource (e.g. "downloaded", "favourited", "rated") which is designated by audience, subject or education level [16]. In contrast to the other usage data formats presented so far, this format is not event, but object-centric. Each data object has exactly one NSDL Paradata record, which is identified by a *recordId* and must contain the URL of the resource to which the paradata record applies (*usageDataResourceURL*). The most important element is the *usageDataSummary*, which comprises all available usage statistics/information about a resource using five different types of values. An *Integer*/*Float* value represents the number of times certain actions have been performed on the resource, e.g. how often it was viewed or downloaded. A *String* value is a textual value that has been associated to the resource, e.g. a comment. A *RatingType* value is the numerical average that represents the judging of a resource on a numerical scale, e.g. a rating according to its usability. A *VoteType* value represents the number of positive and negative responses to a resource, e.g. good or bad for use in classroom. A *RankType* value represents the standing of a resource in a hierarchy, e.g. best of 2010.

Besides its type and value, each *usageDataSummary* element contains the beginning and ending date for the usage data (*dateTimeStart*, *endTimeStart*), information about the *audience* that conducted the event (“educator”, “student”, “general public”, the *subject* of the used resource (e.g. “computing” or “mathematics”) and in which educational level (*edLevel*) the resource was used (e.g. “MiddleSchool”, “Grade 7”).

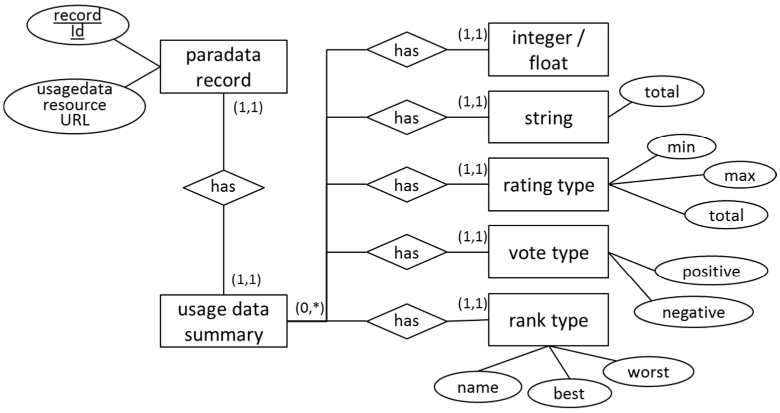

Fig. 4. Simplified excerpt of the NSDL Paradata scheme

For lack of space, these elements are not shown in Fig. 4, but only the elements that are dependent on the type of the *usageDataSummary* element. Please see http://ns.nsdl.org/ncs/comm_para/1.00/records/planets.xml for an extensive example.

3 Conclusion

We reviewed the four most popular usage data representation formats that are being used in the learning domain in this paper and described their main properties. Each format has been created with a specific purpose in mind, so one must be clear about the further applications that will use the collected usage data when choosing the most suitable format.

In order to enhance the interoperability among usage data analysis tools and usage data storage silos, our next step will be to provide guidelines on how mappings between formats can be implemented and what has to be considered. All formats are open and allow supplemental, not pre-defined elements. Additionally, the specified vocabularies are not perceived as complete and for instance in a CAM instance, an *entity* can be described by any metadata scheme. Thus, no one-size-fits-all mapping among the formats is possible. In contrast, mapping can only be defined for specific application scenarios. By providing automatic mapping rules based on specific application scenarios, access to usage data collections will be facilitated. Nevertheless, further work remains to be done in terms of further generalizing the mapping rules so that the automatic conversion tools become less application scenario dependent.

References

1. Schuff, D., Turetken, O., D'Arcy, J., and Croson, D. 2007. 'Managing E-Mail Overload: Solutions and Future Challenges', *Computer*, 40 (2), 31-36.
2. Hauser, J. R., Urban, G. L. Liberali, G., and Braun, M. 2008 'Website Morphing', *Marketing Science*.
3. Adomavicius, G., Tuzhilin, A. 2005. 'Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions', *IEEE Transactions on Knowledge and Data Engineering*, 17 (6), 734-749.
4. Elias, T. 2011. 'Learning Analytics - Definitions, Processes and Potential', retrieved July 4, 2012 from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf
5. Dawson, S. 2011. 'Analytics to Literacies: Emergent Learning Analytics to evaluate new literacies'. *Workshop on New Media, New Literacies, and New Forms of Learning*, London, December 2011. Retrieved July 4, 2012 from http://blogs.ubc.ca/newliteracies/files/2011/12/Dawson.pdf
6. Ferguson, R. 2012. 'The State of Learning Analytics in 2012: A Review and Future Challenges'. *Technical Report KMI-12-01*, Knowledge Media Institute, The Open University, UK. Retrieved July 4, 2012 from http://kmi.open.ac.uk/publications/techreport/kmi-12-01
7. Scheffel, M., Niemann, K., Leony, D., Pardo, A., Schmitz, H.-C., Wolpers, M., Delgado Kloos, C. 2012. "Key Action Extraction for Learning Analytics". Proceedings 7th European Conference on Technology Enhanced Learning (EC-TEL 2012), Springer.
8. Katrien Verbert, Nikos Manouselis, Xavier Ochoa, Martin Wolpers, Hendrik Drachsler, Ivana Bosnic, Erik Duval. 2012. "Context-Aware Recommender Systems for Learning: A Survey and Future Challenges," IEEE Transactions on Learning Technologies, vol. 99, no. PrePrints, , 2012
9. Schmitz, H.-C., Wolpers, M., Kirschenmann, U., Niemann, K. 2012. ‚Contextualized Attention Metadata'. *Human Attention in Digital Environments*, Eds: Claudia Roda, Cambridge University Press, Cambridge, US, 2012 http://www.cup.es/catalogue/catalogue.asp?isbn=9780521765657
10. Çelik, T. 2005. 'Attention.xml Technology Overview', retrieved July 4, 2012 from http://tantek.com/presentations/2005/01/attentionxml.html.
11. Snell, J., Atkins, M., Norris, W., Messina, C., Wilkinson, M., Dolin, R. 2012. JSON Activity Streams 1.0, retrieved July 4, 2012 from http://activitystrea.ms/specs/json/1.0/
12. Snell, J., Atkins, M., Recordon, D., Messina, C., Keller, M., Steinberg, A., Dolin, R. 2012. Activity Base Schema (Draft), retrieved July 4, 2012 from http://activitystrea.ms/specs/json/schema/activity-schema.html
13. Paradata Specification V1.0. Retrieved July 4, 2012 from https://docs.google.com/document/d/1IrOYXd3S0FUwNozaEG5tM7Ki4_AZPrBn-pbyVUz-Bh0/edit?hl=en_US&pli=1#
14. Paradata Cookbook, V 1.0 © Copyright 2011, Learning Registry. CC-BY-3.0. Retrieved July 4, 2012 from https://docs.google.com/document/d/1lggCnowWsDgQxrNjYRAgh2KNwKfq-MV8vLJzRXbAaos/edit?pli=1#
15. Paradata in 20 Minutes or Less, V 1.1 © Copyright 2011, US Department of Education: CC-BY-3.0. Retrieved July 4, 2012 from https://docs.google.com/document/d/1QG0lAmJ0ztHJq5DbiTGQj9DnQ8hP0Co0x0fB1QmoBco/edit?hl=en_US&pli=1#
16. NSDL Paradata, 2012. Retrieved July 4, 2012 from https://wiki.ucar.edu/display/nsdldocs/comm_para+%28paradata+-+usage+data%29