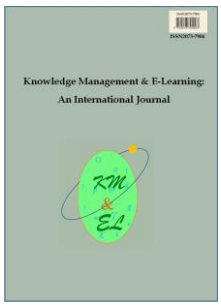
Knowledge Management & E-Learning, Vol.9, No.2. Jun 2017

A generic model for the context-aware representation and federation of educational datasets: Experience from the dataTEL challenge

Julien Broisin
Philippe Vidal
University of Toulouse, France



Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:
Broisin, J., & Vidal, P. (2017). A generic model for the context-aware representation and federation of educational datasets: Experience from the dataTEL challenge. *Knowledge Management & E-Learning, 9*(2), 143–159.

Knowledge Management & E-Learning, 9(2), 143–159

A generic model for the context-aware representation and federation of educational datasets: Experience from the dataTEL challenge

Julien Broisin*

Institut de Recherche en Informatique de Toulouse
University of Toulouse, France
E-mail: julien.broisin@irit.fr

Philippe Vidal

Institut de Recherche en Informatique de Toulouse
University of Toulouse, France
E-mail: philippe.vidal@irit.fr

*Corresponding author

Abstract: Research on online interactions during a learning situation to better understand users' practices and to provide them with quality-oriented features, resources and services is attracting a large community. As a result, the interest for sharing educational data sets that translate the interactions of users with e-learning systems has become a hot topic today. However, the current systems aggregating social and usage data about their users suffer from a series of weaknesses. In particular, they lack a common information model that would allow for exchanges of interaction data at a large scale. To tackle this issue, we propose in this paper a generic model able to federate heterogeneous context metadata and to facilitate their share and reuse. This framework has been successfully applied to several data sets provided by the research community, and thus gives access to a big data set that could help researchers to increase efficiency of existing learning analytics technics, and promote research and development of new algorithms and services on top of these data.

Keywords: Knowledge modelling; Context metadata; Knowledge management; Learning analytics

Biographical notes: Dr. Julien Broisin is an Associate Professor of computer science at the University of Toulouse (France). His research interests include personalized and adaptive learning, inquiry learning through the design and development of remote laboratories, as well as participatory learning through audience response systems.

Pr. Philippe Vidal is a full professor of computer science at the University of Toulouse (France). He leaded the Computer Science Department-Toulouse Institute of Technology for six years before co-leading the Toulouse doctoral school of mathematics, computer science and telecommunications from 2013 to 2014.

1. Introduction

Interest in observation, instrumentation, and evaluation of online educational systems has become more and more important within the Technology Enhanced Learning (TEL) community in the last few years. Conception and development of Adaptive Learning Environments (ALE) in order to classify users, to help and support the creation of recommender systems and intelligent tutoring systems represent a major concern today (Romero, Ventura, Espejo, & Hervas, 2008; Ferguson, 2012).

All these systems ground their adaptation logic on data reflecting interactions of users with electronic information. These data refer to social metadata as well as usage data. Social metadata result from intentional contributions of users and include information like comments, tags, ratings, bookmarks, discussions, reviews, etc. Usage data are automatically collected by the system in the background and reveal relevant interactions between users and electronic artefacts; these usage data are often referred to as paradata and include integration of learning objects into a repository, removal of an activity within an online course, submission of an assignment, and so forth. In this paper, both social metadata and paradata are referred to as context metadata; this perspective on context metadata does not consider content metadata which rely on characteristics and attributes of an electronic resource (e.g., the Learning Object Metadata). Rather, we clearly distinguish raw data that often require further processing before it can be used for adaptation purposes, and inferred data (i.e., indicators) that are derived from transformations, aggregations and other processes operated on the raw metadata.

While context metadata gathered from the adaptive system itself are a good source of implicit feedback, additional data gathered from other sources are meant to improve the adaptation algorithms. Indeed, according to Schafer, Frankowski, Herlocker, and Sen (2007), TEL algorithms are more efficient when: (1) there are many items, (2) there are many users, (3) there are many actions per item, (4) there are more user actions than items to be recommended, (5) users interact with multiple items. Hence, we present in this paper a generic approach to federate heterogeneous context metadata that can be used for adaptation purposes. On one hand, heterogeneity refers to the wide variety of existing learning systems/resources that users are used to deal with, and on the other hand to the unlimited types of context metadata that may be collected. The information model we introduce aims at reaching the following objectives: (1) to be as comprehensive as possible, so that context metadata become meaningful and usable for teachers and for systems as well, (2) to be as flexible as possible, so that diverse adaptation technics can be processed on the basis of a big amount of context metadata collected from any learning artefact.

The paper is organized as follows. Section 2 gives an overview of the adaptation process from our point of view, and exposes some existing approaches focusing on the representation of context metadata to highlight some weaknesses. Section 3 introduces our generic models able to represent both social data and paradata, at both the raw and inferred levels; these models are supported by a set of services that facilitate learning analytics and data mining by learning actors and systems. Section 4 validates our approach by federating several heterogeneous data sets and shows how the resulting data set can be reused and analyzed for various purposes. In Section 5 we discuss some further challenges, while conclusions and future work are provided at the end of the paper.

2. Motivations of this work

Our vision of adaptation is illustrated on Fig. 1 and consists in a loop composed of three distinct phases: (1) the collect of context metadata through dedicated sensors in order to build the knowledge representing the state of the learning situation to be adapted, (2) the data analysis in order to find out adaptation actions to apply, and (3) the execution of the adaptation actions on the learning situation. Besides, this loop can follow two different paths: the second and third phases can be processed either manually or automatically.

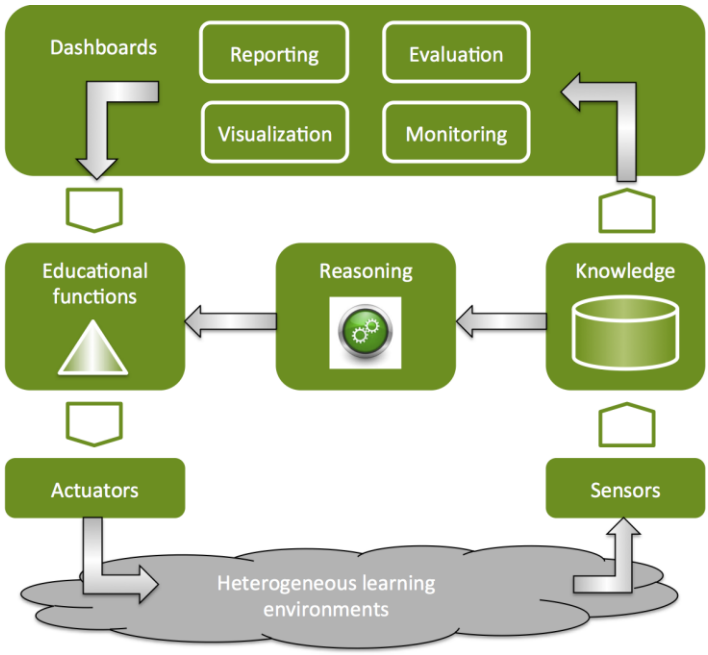


Fig. 1. Adaptation of learning environments

Manual adaptation is handled by users that adapt their learning activities according to various indicators provided by dedicated dashboards and learning analytics technics. Various systems offer teachers and learners diverse dashboards through which actors visualize the learning process and engage manual adaptation actions such as personalization, re-engineering or recommendation activities (Ferguson & Shum, 2012; Mikroyannidis, Gomez-Goiri, Domingue, Tranoris, Pareit, Gerwen, & Marquez-Barja, 2015). These systems perform generally well, since they are designed for a specific situation and expose to users the exact information they need to be able to make the appropriate decision(s) in a given learning situation.

On the other hand, autonomous adaptation consists in continuously analyzing user activities to infer the needs of each student at any moment, and then in applying some of the previous adaptation functions through actuators. To ensure these tasks, some specific modules are required:

- The learner model depicts the characteristics of the learner. Two types of information are represented here: (1) domain independent data (i.e., demographic, previous background, learning style, interests, goals), and (2) domain dependent information which represents the knowledge level of the learner regarding the topics to be studied;

- The content model represents a knowledge structure that describes the concepts related to the domain to be learned. This model may also contain a source of learning material that matches with the domain concepts;
- The tutoring model represents the adaptive engine, and thus integrates some data mining and learning analytics technics such as structured information retrieval, clustering or classification. It computes the learner and content models to reveal what can be adapted, as well as when and how adaptation must be achieved.

The learner model thus acts as a key component of autonomous adaptation (and even manual adaptation, since it is at the basis of the visualization tools provided to users), because adaptive engines make their decisions according to the information available within this model; wrong decisions might be taken if the learner model does not reflect the accurate user experience. The learner model is represented as the Knowledge on Fig. 1. It does not include the learner profile (e.g., the Learner Information Package) only, it also depicts the current and past experiences of the user (Magoulas, Papanikolaou, & Grigoriadou, 2003): it represents the context metadata as defined in Section 1. This model must thus provide as much as possible comprehensive information describing learning experiences, while being as flexible and extensible as possible in order to integrate and to make available a big amount of disparate context metadata. In addition, it should include the indicators that make sense from the educational point of view.

Several initiatives try to provide such a learner model. Based on the Contextualized Attention Metadata (CAM) initiative (Schmitz, Wolpers, Kirschenmann, & Niemann, 2011), Organic.Edunet, a portal offering access to learning resources about agriculture, set up a learner model that focuses on social metadata only (Manouselis & Vuorikari, 2009); it is not possible, for instance, to extend the schema to store usage information other than tags, reviews and ratings. The Learning Registry (Bienkowski, Brecht, & Klo, 2012) is an infrastructure that enables instructors, teachers, trainees and students to discover and use the learning resources held by various American federal agencies and international partners. Learning Registry stores more than traditional descriptive data (metadata) for a learning resource, including social data and paradata that are further shared in a common pool for aggregation, amplification and analysis. However, this framework is application-bounded, being tightly coupled to the learning object concept. Another example is NSDL Paradata (Niemann, Wolpers, Stoitsis, Chinis, & Manouselis, 2013) which aims at providing the educational community with STEM-oriented digital content. This framework collects social metadata restricted to annotation data (i.e., tags and ratings), and stores information about the usage of a digital object in an aggregated way only, thus preventing creation of personalized adaptation process based, for example, on the history of a given user.

Our proposal to enhance existing approaches is introduced in the next section, and stands on a common information model offering a unified view of the various and disparate artifacts composing the user experience.

3. The generic models

Our common information model stands on two generic models characterized by a high level of abstraction. The first model represents raw metadata resulting straight from interactions of users with systems. The other model focuses on inferred data, or indicators, that are calculated after a series of transformations over the raw context metadata.

3.1. The raw context model

The raw context model we designed is illustrated on Fig. 2 and allows for the representation of context metadata collected from heterogeneous web-based learning environments. It is composed of three submodels (i.e., the user context, the environment context, and the usage context), and comprises a set of classes, associations and properties providing a basis for describing diverse artifacts according to more specific learning objectives.

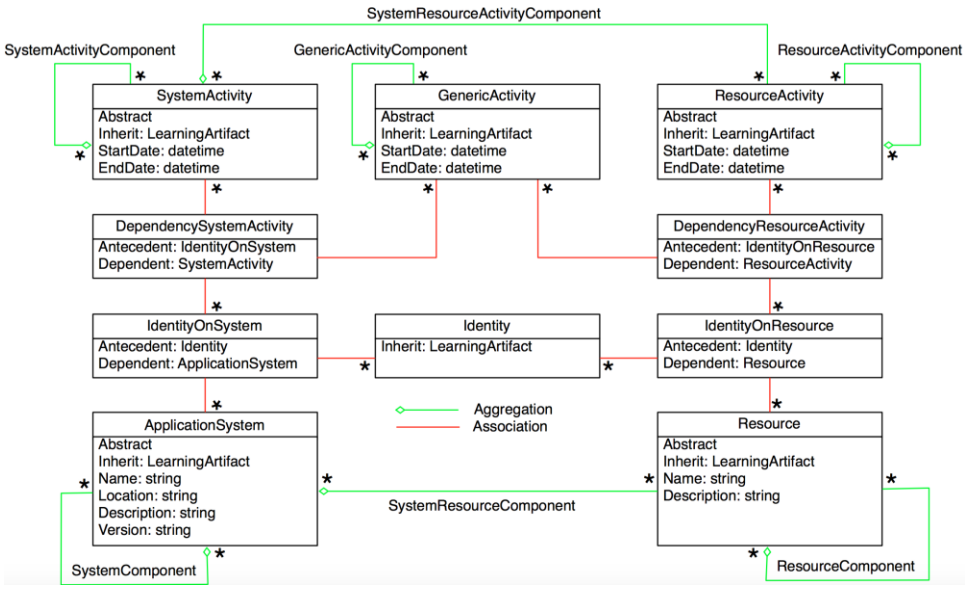


Fig. 2. The generic raw context model

The user context is detailed on Fig. 3. The class *Identity* identifies a user and represents the basis for describing a user. It is characterized by some *PersonalInformation* related to general information about the user such as first name, last name, e-mail, country or birth date. Further, an *Identity* may be described according to its role in a given learning situation; indeed, it is not rare that a user participates in a given course as a teacher, while being a learner in another situation. The abstraction *ProfileCore* represents the top-level class to design any profile specific to TEL actors (e.g., learners, teachers). This class ensures extensibility and openness, and covers any profile that may be required to optimize any TEL application or system. Until now we focused on the learner profile only, represented by the class *LearnerCore* on Fig. 3 and detailed by three subprofiles. The *Cognitive* profile measures learner competencies, tasks and learning styles, the *Knowledge* profile contains information about the actual knowledge levels of a user regarding the concepts of a given ontology, and the *Preference* profile details information about his/her general interests, goals or preferred languages.

The environment context comprises information about the set of electronic artifacts which have been in the focus of the users at any moment. The main classes of the environment model are *ApplicationSystem* and *Resource*; they respectively model any system and resource. Since these systems/resources can be composed of others systems/resources, we introduced two composition relations (i.e., *SystemComponent* and *ResourceComponent* respectively). In addition, another composition (i.e.,

SystemResourceComponent) expresses the fact that a system hosts resources. Finally, in order to link a user with a system or resource, we designed the associations *IdentityOnSystem* and *IdentityOnResource* respectively.

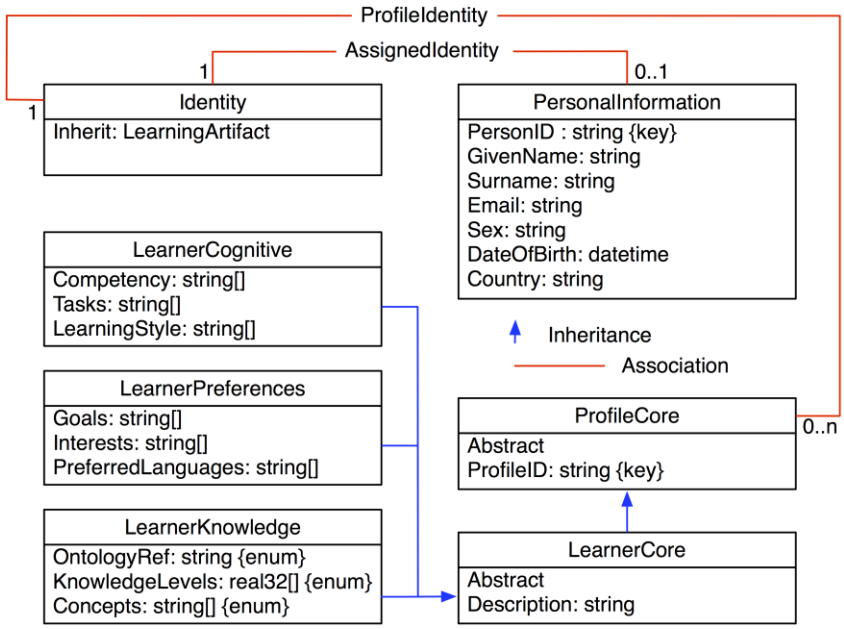


Fig. 3. The generic user model

The usage context contains information describing how users interacted with the environment context. Besides the type of actions performed by users (e.g., search, view, download, etc.), the time when the learning artifact was in the focus of the user, or the duration of the attention, are exposed in the usage context as well. This is composed of three main classes: *ResourceActivity* describes activities specific to learning resources, *SystemActivity* is dedicated to activities operated on learning systems, and *GenericActivity* relies on actions that can be executed on both resources and systems. The aggregation of system/resource activities is possible through the class *SystemActivityComponent* and *ResourceActivityComponent*, while the aggregation of resource activities into system activities is expressed through the class *SystemResourceActivityComponent*. The detailed model can be found in (Butoianu, 2013).

The usage context is connected to the user and environment models through two associations: *DependencyResourceActivity* and *DependencySystemActivity*. The former associates an *IdentityOnResource* (i.e., a tuple <user><resource>) with a *ResourceActivity* to create a tuple <user><resource><activity>; the same reasoning applies to *DependencySystemActivity* to create a tuple <user><system><activity>. By exploiting these associations, various information is made available: the whole set of activities performed by a given user on a specific learning system/resource, or the set of systems/resources on which a given user performed a specific activity, or the users who performed a specific activity on a given learning system/resource.

The resulting raw model tries to reach a good genericity-usability compromise to offer a unified view of heterogeneous context metadata. This generic model isn't application-bounded, as various tools and systems can be represented, but it's not fully general either, thanks to various constraints such as a fixed structure of the root elements

(classes presented in this section can be extended but cannot be modified) and to predefined data types. It's highly expressive too, thanks to various associations and aggregations between the user, environment and usage contexts.

3.2. The indicator model

The generic model presented above is specific to raw contextual data, and thus not well adapted for inspection and interpretation by learning actors and systems; instead, concrete information is needed to monitor and reflect as accurately as possible the progress of the learning activity, and to facilitate data mining and content analytics. Indicators provide a simplified representation of the state of a complex system that can be understood without much training (Glahn, Specht, & Koper, 2007). In the TEL area, indicators may be of different nature, depending on the learning goals, actions, performances, outcomes as well as the situation in which the learning process takes place (Florian, Glahn, Drachsler, Specht, & Gesa, 2011). Therefore, we designed a generic indicator model characterized by the main following properties: it distinguishes clearly indicator definition and indicator value, and may describe any artifact of the raw context model.

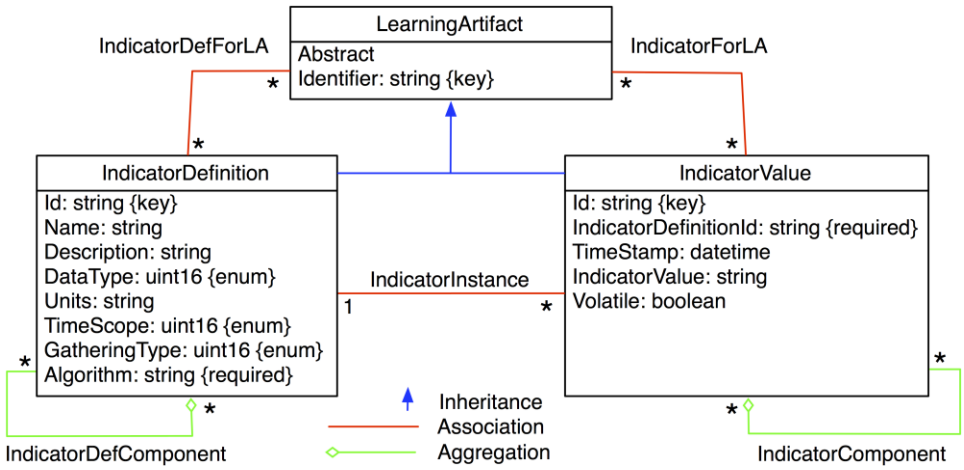


Fig. 4. The generic indicator model

The resulting model is illustrated on Fig. 4 and is composed of two main classes. The class *IndicatorDefinition* behaves as a pattern that specifies the semantics and usage of an indicator (i.e., its metadata), it does not capture the value of the indicator (the class *IndicatorValue* holds this information). Additional metadata for an indicator can be provided by subclassing the class *IndicatorDefinition*, but the most important descriptors are *Name* (i.e., a human readable name of the indicator), *Description* (i.e., a human readable description of the objective of the indicator), *DataType* (i.e., the data type of the indicator; for example, "boolean", "datetime", "integer" or "string" may be specified), *Units* (i.e., the specific units of the indicator; examples are actions, second), *TimeScope* (i.e., the time scope to which the indicator value applies), *GatheringType* (i.e., the way the indicator value is calculated; examples are "periodically", "on request", or at the time the indicator definition is "created"), and *Algorithm* (i.e., the algorithm leading to the calculation of the indicator value by the underlying instrumentation). In addition, the composition relation *IndicatorDefComponent* makes it possible to reuse indicators in order to define high-level indicators standing on the definition of lower-level indicators.

The class *IndicatorValue* acts as a container of values. A single value is stored in each instance of this class, and each of the instances is associated with an indicator definition. The main properties of this class are *TimeStamp* to indicate the time when the value has been computed, *IndicatorValue* which is the value itself, stored as a string, and *Volatile* which specifies if a new instance must be created when a new value is calculated, or if the existing instance must be updated.

In addition, the generic indicator model defines several associations to interlink learning artifacts, indicator definitions, and indicator values:

- *IndicatorDefForLA* specifies the definitions that apply to a given learning artifact. A specific definition may apply to any artifact of the raw context model, and a given artifact can be characterized by an unlimited number of indicator definitions.
- *IndicatorForLA* links indicator values to learning artifacts. Here again, a single value may apply to one or several learning artifacts, and a given entity can be characterized by an unlimited number of indicator values.
- *IndicatorInstance* links an indicator value to its definition. A value applies to a single definition, but a definition may be linked to several values.

The generic indicator model suggested here gives the opportunity to express statistical and arithmetical indicators, but also to define a wide variety of more or less complex indicators. The clear distinction between indicators' definition and value brings several advantages, especially regarding their reuse. On one hand, the metadata describing the definition of an indicator makes it easy for designers of dashboards (in case of manual adaptation) or reasoning modules (in case of autonomous adaptation) to identify precisely the nature and objective of the inferred data so it can be easily integrated into the adaptive process. On the other hand, designers of adaptive frameworks can easily apply an existing indicator to an artifact specific to their learning situation (e.g., if an indicator has been defined to reveal the number of activities performed on a given learning resource, the same definition can be used to retrieve the number of activities that have been operated on a given learning system); in addition, as described in the next section, they don't have to consider the way it is calculated and can thus focus on their primary tasks (i.e., visualization and processing). Finally, the indicator model allows assigning several values to the same definition, thus offering the opportunity to retrieve the history of a given indicator, that is the user experience history.

In this section we designed a generic information model to represent heterogeneous context metadata. It is characterized by a structured representation that makes it easy to find relevant data effectively and to avoid duplication of data, and provides extensibility required to collect information of future applications. The raw context model allows expressing statements such as "This user did this with this entity", where "this user" represents any learning actor, "did this" comprises any type of social and usage activities, and "this entity" refers to any electronic artifact. Since indicators are based on the wide variety of context metadata that can be described through this generic model, richer data can be inferred. These data come to supplement the user experience based on the raw model by providing very comprehensive and meaningful data.

In the context of Computer-Supported Collaborative Learning (CSCL), Harrer, Martínez-Monés, and Dimitracopoulou (2009) designed a joint format that could be used by the analysis tools of the Kaleidoscope consortium in order to support students and teachers during online learning activities in a collaborative setting. The common format they propose is in line with the generic models exposed in this section, as it allows to

track user interactions based on the same paradigm: "at least one user did this activity, eventually with this object". This format also stands on a core structure that can be extended by defining "additional information"; however, this field, combined with the XML-like representation of the basic structure, lacks semantics to explicitly and precisely express new data to collect. The broader objective of this common format is to foster adoption of interactions analysis tools by the CSCL community (Martínez-Monés, Harrer, & Dimitriadis, 2011); concerning this point, the common format lacks the possibility to specify inferred data. Currently, indicators designed by teachers (i.e., data meant to have a significant pedagogical added value) within a given interaction analysis tool, and based on the data collected according to the common format, cannot be easily shared with the community and reused within others tools.

The next section exposes some extensions of the generic model that meet the specificities of diverse learning situations, and then explores a data set resulting from the federation of social and usage data to show how it can be used for adaptation purposes.

4. Case-study: A federation of TEL data sets

The dataTEL challenge was launched as part of the first workshop on Recommender Systems for TEL (Manouselis, Drachsler, Verbert, & Santos, 2010), jointly organized by the 4th ACM Conference on Recommender Systems and the 5th European Conference on Technology Enhanced Learning in September 2010. This call invited research groups to submit existing data sets from TEL applications that can be used for research purposes. To date, ten (10) providers detailed in (Verbert, Drachsler, Manouselis, Wolpers, Vuorikari, & Duval, 2011) submitted a proposal. These include: Mendeley (Jack, Hammerton, Harvey, Hoyt, Reichelt, & Henning, 2010), a research portal that helps users to organize their research, collaborate with colleagues and discover new knowledge; APOSDLE (Ghidini, Pammer, Scheir, Serafini, & Lindstaedt, 2007), a Personal Learning Environment (PLE) that leverages the productivity of workers by integrating learning within everyday work task; ReMashed (Drachsler, Rutledge, van Rosmalen, Hummel, Pecceu, Arts, Hutten, & Koper, 2010), a recommender web portal that aggregates contributions from a variety of web 2.0 services such as delicious, youtube, or flickr; Organic.Edunet (Manouselis & Vuorikari, 2009), MACE (Wolpers, Memmel, & Giretti, 2009), Travel well (Vuorikari & Van Assche, 2007) and CGIAR (Zschocke, Beniest, Paisley, Najjar, & Duval, 2009), some web portals that federate various learning object repositories; ROLE (Santos, Verbert, Govaerts, & Duval, 2011), a platform that enables learners to build their own PLE through the assembly of various widgets; SidWeb (Ochoa, Ternier, Parra, & Duval, 2006), a LMS used at the Escuela Superior Politecnica del Litoral, Ecuador; UC3M (Romero-Zaldivar, Pardo, Burgos, & Delgado Kloos, 2012), a LMS that collects data from a virtual machine used in a C programming course. In addition, usage data collected from the Moodle server deployed within our university (Moodle UT) are included in this study as well.

The objective of this case-study is twofold: first, to show how the modeling approach exposed in the previous section can be successfully applied to federate data stemming from the above learning systems; second, to provide researchers with a big collection of data to compare the results of different adaptation algorithms and the influence of context metadata on the adaptation process.

4.1. Extension of the generic models

Our methodology to identify the extensions required to federate the dataTEL data sets consisted in three steps: the analysis of each context metadata collected by each system, the aggregation of data characterized by common properties, and the design of the extended models. Notice that the studied data sets, except Moodle UT and Travel well, do not provide information about a user but an identifier, thus the user model depicted on Fig. 3 has been reused without modifications (our user context takes into account the Moodle fields that are useful for adaptation purposes, together with the country, language and interests provided by Travel well). Also, for readability reasons, the extended environment and usage contexts are not illustrated by figures; Table 1 gives a synthetic view about the applications, resources and activities considered by the federated data set.

4.1.1. Extension of the environment context

We identified fifteen (15) different applications and tools observed within the dataTEL data sets, classified as *Desktop* or *WebApplications*. Indeed, even if the dataTEL providers imply three different kinds of applications only (i.e., *LearningManagementSystem*, *PersonalLearningEnvironment* and *WebPortal*), some of them collect data from other sources as well: ROLE collects information about interactions of users with various *Widgets*, a *ChatApplication* and a *KnowledgeMap*; SidWeb captures users activities from a discussion *Forum* and a *Quiz* tool; MACE, Travel well and CGIAR monitor users interactions with multiple *LearningObjectRepositories*; UC3M tracks users interacting with various applications such as a *WebBrowser*, a command line *Interpreter*, a *TextEditor*, a *MemoryProfiler*, a C *Compiler*, and a C *Debugger*. In addition, we extended the aggregation relation between learning systems (see Fig. 2) to express the fact that a web application can be composed of one or more widgets.

Seventeen (17) types of resources are currently listed within the dateTEL providers. Most of the systems collects information about learning objects, as defined by the IEEE LOM P1484.12 working group; we thus defined a *LearningObject* as an abstract artifact that may refer to an *Article*, a *WebPage*, a *Simulation*, a *Presentation*, an *Assignment*, a *Submission*, a *Quiz*, a *Message* or a *File*. In addition, some systems denote the aggregation of learning objects into *Collections* such as *BookmarkList*, *Courseware*, *ArticleCollection*, *ChatRoom* or *DiscussionThread*; we designed the matching aggregations by extending the *ResourceComponent* relation. The other types of resources are *ShellCommand* executed through an interpreter, and *Ontology* and *Topic* that are required by Aposdle to monitor navigation of users through these resources.

Finally, the aggregation relation *SystemResourceComponent* between systems and resources has been intensively extended to translate various statements: LMS/PLE host courseware, interpreters execute commands, learning object repositories store learning objects, and so forth.

4.1.2. Extension of the usage context

The extension of the generic usage model is meant to express how and when the learning artifacts described above have been used by users. The methodology consisted here in the identification of the activities that apply to both applications and resources (e.g., *Open*, *Close*, *Rate*, *Review*, *Tag*, *Search*, *Download*), and then in the recognition of the activities specific to applications, as well as those specific to resources. Both sets of activities have

been analyzed to identify the generic application activities (e.g., *LogIn*, *LogOut*, *Install*, *Uninstall*) and the generic resource activities (e.g., *Edit*, *Create*, *Delete*). Finally, for each application and resource, the matching specific activities have been specified. No activities specific to an application of the dataTEL data sets was found. However, some activities specific to learning objects (e.g., *IndexIntoLor*, *RemoveFromLor*, *AddToCourseware*, *RemoveFromCourseware*, *AddToCollection*), messages (e.g., *Post*, *SendTo*), commands (e.g., *Execute*), quizzes (e.g., *StartAttempt*, *FinishAttempt*) and topics (e.g., *Perform*, *UpdateExperienceLevel*) have been designed.

The extensions of the generic models lead to an information model federating the data observed by each data set of the dataTEL initiative. Thanks to the aggregation and association relations, our modeling approach focuses on semantically quality-oriented context metadata by contextualizing accurately and intelligibly users interactions with learning applications and resources. As an illustration, context metadata translating the consultation of a web page differ according to the system giving access to that web page: in case of Moodle UT, the context metadata refer, in addition to the user and the web page itself, to the course integrating the web page and to the LMS hosting the courseware; in case of a web portal such as MACE, the context metadata refer to the repository storing the link to that web page as a learning object.

4.2. Integration of the dataTEL datasets

To support the generic and extended models, we designed an infrastructure standing on two main proposals: a repository ensuring consistency of context metadata but also able to manage indicators, and a set of modules built on top of the repository to facilitate data management. Following our object-oriented approach, the repository implements the Oracle Object Relational Database that combines the advantages of both relational and object-oriented paradigms: data are modeled as objects (thus offering a one-to-one mapping of our models without semantics losses) but stored into tables and manipulated through the SQL query language; compared to XML-oriented databases such as eXist, the solution we adopted responds much faster to complex queries (Butoianu, 2013). To manage indicators, we designed three distinct modules as callback handler procedures: the event manager monitors some specific events occurring inside the repository, the indicator handler is responsible for the calculation of indicator values, and the indicator notifier offers the opportunity to execute actions outside the repository. When a new indicator definition (or a new association between an existing indicator and a learning artifact) is created, the event manager notifies the indicator handler so that the value(s) is(are) calculated according to the matching definition and, once a new value is available, the event manager alerts the indicator notifier so that external actions can occur.

A set of modules has been designed to facilitate the communication with the repository. This toolbox, developed as web services, currently comprises three main services: one to index new context metadata, another to retrieve existing context metadata, and a third service to subscribe to indicators. While the first two services are independent from any query language and result format (thanks to the Simple Query/Publishing Interfaces), the latter stands on the publish/subscribe paradigm to promote reuse of indicators: any system or application can subscribe to indicators of interest and receive notifications as soon as a new value is calculated within the repository (first, the notifier module sends the new value to the indicator service which, in its turn, notifies the subscribers). Indicators values are thus delivered, at the right time, to any adaptive component interested in the analysis of users' behavior.

Table 1
Statistics about the federated data

	Resources	Activities	Content of context metadata	Number of activities
Moodle UT	Course (1,988)	Create, view, update, delete	User (U)+Course (C)+LMS+A	1,602,667
	WebPage (18,289)	View, download, rate, delete, integrate	U+WebPage(W)+C+LMS+A	1,137,021
	Assignment (2,602)	Download, view, update, delete	U+Assign.+C+LMS+A	334,431
	Submission (1,860)	Create, download	U+Sub.+Assign.+C+LMS+A	126,620
Aposdle	File (110)	View	U+File+C+PLE+A	197
	WebPage (11)	View	U+W+C+PLE+A	16
	Topic (192)	Perform, view, update experience level	U+Topic+Ontology+PLE+A	1,452
TW	LearningObject (1,605)	Search, rate, tag	U+LO+LOR+WebPortal (WP)+A	9,563
	Quiz (413)	Rate, tag	U+Quiz+WP+A	1,154
MACE	WebPage (12,369)	Tag, view, rate	U+W+LOR+WP+A	122,605
MD	Article (32,285)	Add to collection, view, rate	U+Article (in collection)+ WP+A	145,461
	71,724 resources	13 types of activities	Total number of activities	3,481,187

The objective of this infrastructure is to build a data warehouse based on the generic and extension models that gives unified access to the whole set of data provided by the dataTEL contributors, e.g., for educational data mining and learning analytics purposes (see section 4.3). Thus, on the basis of this infrastructure, we integrated the dataTEL data sets into the repository. The integration process consisted of the three following steps: (i) to get each data set from the dataTEL contributors (we asked each contributor access to their data); (ii) to design integrators specific to each data set ensuring the mapping between the considered context metadata schema and our models (according to the nature of the data source, e.g., Microsoft Excel, SQL or XML files, the main operations performed on the data by the integrators were extraction of each file record and mapping towards our generic schema); (iii) to store the resulting metadata records of each dataTEL provider into the repository using the indexation service described above. Unfortunately, some dataTEL contributors did not agree to provide us with their data, as the users whose interactions have been recorded did not explicitly and

formally consent to share the data with external parties. Thus, only four (4) of the ten (10) data sets have been made available: Mendeley (MD in Table 1), Aposdle, MACE and Travel well (TW in Table 1); we included the data collected from Moodle UT as well. Statistics about the resulting data set are exposed in Table 1; as six (6) of the ten (10) data sets are missing, several resources and activities of the extended models do not appear in this table. Moreover, none of the data sets collect application-related activities.

4.3. Usefulness of the resulting data set

4.3.1. Assist teachers and learners

In addition to the growing number of richer indicators that could be pushed to teachers and learners through dashboards, the federative data set could promote large-scale community of practices and facilitate collaborative knowledge building and sharing. Indeed, even if curriculums are often replicated from one learning organization to another, community of practices are mostly limited to actors of a local organization (e.g., a university). Since our data set hosts experiences of users that interact with various platforms, a service could be set up to build coherent community according to teachers and learners' educational interests. Actors of a given body would be able to identify peers located in others organizations, and thus to mutualize their experiences in terms of teaching and learning skills.

4.3.2. Improvement of existing algorithms and services

As mentioned in Section 1, Schafer et al. (2007) have stated that collaborative filtering algorithms become more and more efficient as the mass of data available within the system is significant. Both MACE and Travel well provide ratings about learning resources on the same scale (i.e., 1-5), and implement the aforementioned algorithm to recommend content to their users. Using our federated data set as input instead of their respective context records should improve efficiency of the recommendations; an evaluation is being conducted to confirm this hypothesis.

The cold start problem is a well-known issue that prevents the well-functioning of adaptive systems from the very beginning (Lam, Vu, Le, & Duong, 2008). Obviously, access to the large amount of context metadata stored into the repository is meant to tackle this issue. For instance, on the basis of the federated data set, we designed and integrated a recommender system for learning resources within Moodle UT. In addition to exposing resources provided by external systems, the collaborative filtering algorithm we implemented was up and running as soon as this tool was available to users.

4.3.3. Design of new algorithms and services

Research on online interactions in learning situations to better understand users' practices and to provide them with quality-oriented features, resources and services is attracting a large community. On one hand, the federative data set we propose here allows for replications of adaptation algorithms over heterogeneous data: comparative, cumulative and contrastive data mining can be processed to reveal the algorithms that perform best in a given learning situation (Verbert, Manouselis, Drachsler, & Duval, 2012). On the other hand, smart learning environments aim at supporting learners by combining the use of innovative technologies and the adoption of pedagogical approaches that best fit the

learner context. These environments are neither fully technologic nor tightly coupled to a given educational theory, instead they have to establish the most proper compromise between these two aspects by self-adapting to the changes of the user experience. Associated to the relevant algorithms, and according to the learning context and situation, the heterogeneity of the federated data set makes it possible to elaborate various services to build dynamic user-centric learning environments that provide actors with various types of entities.

Some work is in progress to provide users with a smart learning environment, acting as a portal, which exploits the federated data set. Currently, according to the cognitive profile of users (in terms of learning preferences and interests) described in the repository, the system dynamically pushes web pages, learning objects and assignments coming from Moodle UT, Mace and Mendeley. The system adopts various widgets such as those exposed to users in the ROLE context (Santos et al., 2011) to visualize the content of the resources and to ensure communication with the target system. This work is still in an early stage; thus, the outcomes of this environment cannot be discussed at the time of writing this article.

5. Privacy concerns

A major concern that must be addressed by context-aware systems is the user privacy, since they collect, store and process confidential and sensitive data about users.

At first sight, the generic models illustrated on Fig. 2 and Fig. 4 do not hold sensitive data: they represent activities that have been performed by a given identity on a set of resources and applications, but no information is detailed about this identity. Instead, the user profile depicted on Fig. 3 reveals very personal and sensitive data. An intuitive solution to tackle the privacy issue would consist in removing this model from the context metadata repository, but some interesting functionalities and analytics wouldn't be possible anymore (e.g., content-based and collaborative filtering, recommendations of learning paths, etc.). However, considering that the learning profile of a user cannot reveal his/her identity, *PersonalInformation* is the single class to be removed; adaptive systems could retrieve the profile of users using their identifier, and then process the adaptation algorithms of their choice.

At a closer look, one can wonder if advanced and repeated data mining technics over the context repository could not lead to the identification of users. Indeed, a single context metadata record has no significant meaning, but the linkage of an important number of records may have: the more context metadata are collected, the more information about users is detailed, and the higher the chances to identify a user are. Therefore, even if the data sets have been anonymised, the dataTEL providers do not naturally agree to share context metadata at a large scale. The federative data set is thus currently open to the providers' community, but some investigations must be leaded to guarantee the anonymisation of data before opening the repository worldwide.

6. Conclusions

We have presented in this paper a comprehensive generic model able to offer a unified view of context metadata collected from heterogeneous learning tools and resources. This representation of the user experience is: structured to facilitate relevant and efficient filtering and crosswalk across data, extensible to integrate current and future context-

aware applications, and uniform to facilitate the interpretation of data during the adaptation process. The generic model has been extended to federate the data sets of the dataTEL challenge. The resulting data are held at the disposal of this community and can be used as a basis for further analytics to leverage adaptation algorithms and services.

Besides the dataTEL challenge, others initiatives encourage researchers to share their data sets. We are particularly interested in the DataSHOP initiative that provides several data sets collected from adaptive systems such as intelligent tutoring systems. The challenge will be to align these data with the generic model.

Even if our storage infrastructure based on an object-relational database suits perfectly our object-driven approach, some investigations are being leaded to identify some alternatives that would increase performance in terms of execution of complex queries that are required by advanced autonomous adaptation algorithms; the federative data set is intended to become bigger and bigger, thus the scalability issue has to be tackled. In addition, even if the SOAP API contributes to the privacy of the context metadata, it requires more development efforts to be invoked by external partners; a REST API has to be designed to facilitate access to the federative data set at a large scale and to bring our framework into compliance with nowadays web 3.0 tools.

References

Bienkowski, M., Brecht, J., & Klo, J. (2012). The learning registry: Building a foundation for learning resource analytics. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Vancouver, Canada.

Butoianu, V. (2013). *Share and reuse of context metadata resulting from interactions between users and heterogeneous web-based learning environments*. Doctoral thesis, University of Paul Sabatier, Toulouse, France.

Drachsler, H., Rutledge, L., van Rosmalen, P., Hummel, H., Pecceu, D., Arts, T., Hutten, E., & Koper, R. (2010). ReMashed - An usability study of a recommender system for mash-ups for learning. *International Journal of Emerging Technologies in Learning, 5*(SI: ICL2009), 7–11.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317.

Ferguson, R., & Shum, S. B. (2012). Social learning analytics: Five approaches. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Vancouver, Canada.

Florian, B., Glahn, C., Drachsler, H., Specht, M., & Gesa, R. F. (2011). Activity-based learner-models for learner monitoring and recommendations in Moodle. *Lecture Notes in Computer Science, 6964*, 111–124.

Ghidini, C., Pammer, V., Scheir, P., Serafini, L., & Lindstaedt, S. (2007). APOSDLE: Learn@ work with semantic web technology. In *Proceedings of the IMEDIA and I-SEMANTICS*. Graz, Austria.

Glahn, C., Specht, M., & Koper, R. (2007). Smart indicators on learning interactions. In *Proceedings of the 2nd European Conference on Technology Enhanced Learning*. Crete, Greece.

Harrer, A., Martínez-Monés, A., & Dimitracopoulou, A. (2009). Users' data. In N. Balacheff, S. Ludvigsen, T. de Jong, A. Lazonder, & S. Barnes (Eds.), *Technology-Enhanced Learning* (pp. 175–193). Springer Netherlands.

Jack, K., Hammerton, J., Harvey, D., Hoyt, J. J., Reichelt, J., & Henning, V. (2010).

Mendeley's reply to the dataTEL challenge. *Procedia Computer Science, 1*(2), 1–3.

Lam, X. N., Vu, T., Le, T. D., & Duong, A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*. Suwon, Korea.

Magoulas, G. D., Papanikolaou, Y., & Grigoriadou, M. (2003). Adaptive web-based learning: Accommodating individual differences through system's adaptation. *British Journal of Educational Technology, 34*(4), 511–527.

Manouselis, N., Drachsler, H., Verbert, K., & Santos, O. C. (2010). Workshop on recommender systems for technology enhanced learning. In *Proceedings of the 4th ACM Conference on Recommender Systems* (pp. 377–378). New York, USA.

Manouselis, N., & Vuorikari, R. (2009). What if annotations were reusable: A preliminary discussion. *Lecture Notes in Computer Science, 5686*, 255–264.

Martínez-Monés, A., Harrer, A., & Dimitriadis, Y. (2011). An interaction-aware design process for the integration of interaction analysis into mainstream CSCL practices. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL* (pp. 269–291). Springer.

Mikroyannidis, A., Gomez-Goiri, A., Domingue, J., Tranoris, C., Pareit, D., Gerwen, V. V., & Marquez-Barja, J. (2015). Deploying learning analytics for awareness and reflection in online scientific experimentation. In *Proceedings of the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning*. Toledo, Spain.

Niemann, K., Wolpers, M., Stoitsis, G., Chinis, G., & Manouselis, N. (2013). Aggregating social and usage datasets for learning analytics: Data-oriented challenges. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*. Leuven, Belgium.

Ochoa, X., Ternier, S., Parra, G., & Duval, E. (2006). A context-aware service oriented framework for finding, recommending and inserting learning objects. *Lecture Notes in Computer Science, 4227*, 697–702.

Romero, C., Ventura, S., Espejo, P., & Hervas, C. (2008). Data mining algorithms to classify students. In *Proceedings of the 1st International Conference on Educational Data Mining*. Montreal, Quebec, Canada.

Romero-Zaldivar, V. A., Pardo, A., Burgos, D., & Delgado Kloos, C. (2012). Monitoring student progress using virtual appliances: A case study. *Computers & Education, 58*(4), 1058–1067.

Santos, J. L., Verbert, K., Govaerts, S., & Duval, E. (2011). Visualizing PLE usage. In *Proceedings of the 1st Workshop on Exploring the Fitness and Evolvability of Personal Learning Environments*. La Clusaz, France.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. *Lecture Notes in Computer Science, 4321*, 291–324.

Schmitz, H. C., Wolpers, M., Kirschenmann, U., & Niemann, K. (2011). Contextualized attention metadata. In C. Roda (Ed.), *Human Attention in Digital Environments* (pp. 186–209). Cambridge University Press.

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). Banff, AB, Canada.

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology & Society, 15*(3), 133–148.

Vuorikari, R., & Van Assche, F. (2007). Collaborative content enrichment in multilingual Europe, European Schoolnet approach on educational resources. In *Proceedings of the 3rd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*. Barcelona, Spain.

Wolpers, M., Memmel, M., & Giretti, A. (2009). Metadata in architecture education-first evaluation results of the mace system. *Lecture Notes in Computer Science, 5794*, 112–126.

Zschocke, T., Beniest, J., Paisley, C., Najjar, J., & Duval, E. (2009). The LOM application profile for agricultural learning resources of the CGIAR. *International Journal of Metadata, Semantics and Ontologies, 4*(1/2), 13–23.