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.

Dataset-Driven Research to Support Learning and Knowledge Analytics

Katrien Verbert[1], Nikos Manouselis[2], Hendrik Drachsler[3] and Erik Duval[1]

[1]Dept. Computerwetenschappen, KU Leuven, Belgium // [2]Agro-Know Technologies, Athens, Greece & University of Alcala, Spain // [3]CELSTEC, Open University of the Netherlands, The Netherlands //
katrien.verbert@cs.kuleuven.be / /nikosm@ieee.org // hendrik.drachsler@ou.nl // erik.duval@cs.kuleuven.be

ABSTRACT

In various research areas, the availability of open datasets is considered as key for research and application purposes. These datasets are used as benchmarks to develop new algorithms and to compare them to other algorithms in given settings. Finding such available datasets for experimentation can be a challenging task in technology enhanced learning, as there are various sources of data that have not been identified and documented exhaustively. In this paper, we provide such an analysis of datasets that can be used for research on learning and knowledge analytics. First, we present a framework for the analysis of educational datasets. Then, we analyze existing datasets along the dimensions of this framework and outline future challenges for the collection and sharing of educational datasets.

Keywords

Learning and knowledge analytics, Datasets, Open science

Introduction

The need for better measurement, collection, analysis and reporting of data about learners has been identified by several researchers in the Technology Enhanced Learning (TEL) field (Siemens 2010; Romero et al. 2007; Duval 2011). This need has been translated into an emerging strand of research on learning and knowledge analytics (LAK), as reflected by a number of conferences and special issues in recent years (Siemens & Gasevic, 2011). Among others, the analysis of learner data and identification of patterns within these data are researched to predict learning outcomes, to suggest relevant resources and to detect error patterns or affects of learners. These objectives are researched to act upon needs of a variety of stakeholders, including learners, teachers and organizations. This is what drives major initiatives such as the US-based Learning Registry (http://www.learningregistry.org) to collect data and make them publicly available for research and application purposes.

Siemens (2010) defines learning analytics as “the use of intelligent data, learner-produced data, and analysis models to discover information and social connections, and to predict and advise on learning.” Contributions to the first conference on learning analytics and knowledge in 2011 indicate that information visualization, social network analysis and educational data mining techniques offer interesting perspectives for this emerging field. Whereas the specific techniques differ depending on context and the intended goals, the main objective of the approaches is to identify needs of target users and to support these needs using intelligent and adaptive systems.

Despite the recognition of the importance of LAK, the literature related to this topic is rather limited. Research on web analytics, search engines and recommender systems are excellent examples of how data gathering during an analytics cycle can be used to refine offerings to users (Elias, 2011). Whereas several recommender systems for learning (Manouselis et al., 2011), intelligent tutoring systems (Romero et al. 2007) and visual analytics systems (Govaerts et al. 2010) have been implemented for use in learning scenarios in recent years, many of these intelligent tools often stay in researcher hands and rarely go beyond the prototype stage (Reffay & Betbeder, 2009). Among others, researchers have argued that the time needed by social science to validate prototypes is too long compared to the rate of technology innovation.

An important component to facilitate research in this area is the existence of extensive overviews of the available datasets that will provide researchers with a wide array of potential data sources to experiment with, as well as with an analysis of their properties that will help researchers decide about their appropriateness for their experiments. Such an overview is missing from LAK today, since only initial attempts have been made to document and study existing datasets (Drachsler et al., 2010). In this article, we extend our initial analysis (Verbert et al., 2011) of the datasets collected by the dataTEL Theme Team of the European Network of Excellence STELLAR (http://www.teleurope.eu/pg/groups/9405/datatel/), in order to provide a more comprehensive overview of datasets for LAK research. The article makes four primary research contributions:

- First, we present related initiatives that are collecting datasets and the needs and opportunities to make educational datasets available for LAK research.
- Second, we present a framework for the analysis of educational datasets. In particular, we present properties of educational datasets and LAK objectives that can benefit from the availability of such data.
- Third, we analyze existing datasets along the dimensions of our educational dataset framework. We also discuss existing research that has used these datasets for LAK related research.
- Finally, we present future challenges to enable the sharing and reuse of datasets among researchers in this field.

Background

In an increasing number of scientific disciplines, large data collections are emerging as important community resources (Chervenak et al., 2000). These datasets are used as benchmarks to develop new algorithms and compare them to other algorithms in given settings (Manouselis et al., 2010b). In datasets that are used for recommendations algorithms, such data can for instance be explicit (ratings) or implicit (downloads and tags) relevance indicators. These indicators are then for instance used to find users with similar interests as a basis to suggest items to a user.

To collect TEL datasets, the first dataTEL Challenge was launched as part of a workshop on Recommender Systems for TEL (RecSysTEL, Manouselis et al., 2010a) that was jointly organized by the 4th ACM Conference on Recommender Systems and the 5th European Conference on Technology Enhanced Learning in September 2010. In this call, research groups were invited to submit existing datasets from TEL applications. A special dataTEL Cafe event took place during the RecSysTEL workshop in Barcelona to discuss the submitted datasets and to facilitate dataset sharing in the TEL community.

Related work is carried out at the Pittsburgh Science of Learning Center (PSLC). The PSLC DataShop (Stamper et al., 2010) is a data repository that provides access to a large number of educational datasets derived from intelligent tutoring systems. Currently, more than 270 datasets are stored that record 58 million learner actions. Several researchers of the educational data mining community have used these datasets to predict learner performance.

The Mulce project (Reffay & Betbeder, 2009) is also collecting and sharing contextualized interaction data of learners. A platform is available to share, browse and analyze shared datasets. At the time of writing, 34 datasets are available on the portal, including a dataset of the Virtual Math Teams (VMT) project. This project investigated the use of online collaborative environments to support K-12 mathematics learning. These datasets have been used extensively by the Computer Supported Collaborative Learning (CSCL) community (Stahl, 2009).

Other efforts have been driven by fields studying child language acquisition. The CHILDES system (MacWhinney, 1996, 2007) helped realize much advancement in this field through sharing language-learning data. TalkBank (http://talkbank.org) is a follow-up project that is researching guidelines for ethical sharing of data, metadata and infrastructure for identifying available data, and education of researchers to the existence of shared data, tools, standards and best practices.

LinkedEducation.org is another initiative that provides an open platform to promote the use of data for educational purposes. At the time of writing, five organizations have contributed datasets. Available datasets describe the structure of organizations and institutions, the structure of courses, learning resources and interrelationships between people. In addition, various schemas and vocabularies are provided to describe the internal structure of an academic institution, discourse relationships, activity streams in social networks and educational resources. Such schemas and vocabularies offer interesting perspectives for the sharing and reuse of educational interaction data that is relevant for LAK research.

Several other initiatives are available that focus on providing the means to share datasets among researchers in a more generic way. DataCite.org is an organization that enables users to register research datasets and to assign persistent identifiers to them, so that datasets can be handled as citable scientific objects. The Dataverse Network (King, 2007) is an open-source application for publishing, citing and discovering research data. The network was established at Harvard University and is aimed to increase scholarly recognition for data contributions. Fact sheets of datasets are gathered from organizations and researchers are encouraged to make the data publicly available, if possible. The Australian National Data Service (Treloar & Wilkinson, 2008) is a similar initiative in Australia that works on services to help researchers persistently identify and describe data.

In this paper, we analyze educational datasets that have been collected by dataTEL and related initiatives. We focus specifically on datasets that contain interaction/usage data of learners and that can be used for analytics' research. In the next section, we present a framework for educational datasets that can be used to describe and analyze educational datasets. In addition, we discuss how work of related initiatives fits within this framework. Then, we analyze available datasets along the dimensions of this framework.

A Framework for Educational Datasets

In this section, we present a framework for the analysis of educational datasets. The framework is intended to address questions researchers might have about the potential usefulness of a dataset for their research purposes.

As illustrated in Figure 1, the framework constitutes three parts. *Dataset properties* describe the overall dataset, such as the application and the educational setting from which the data was collected. *Data properties* define at a finer grained level where data elements available, including action types such as downloads or selects and information about the learner and other entities involved. The third part of the framework defines a list of *objectives* of LAK research. These objectives are mapped to dataset and data properties in the next section to determine the potential usefulness of a dataset for LAK research purposes.

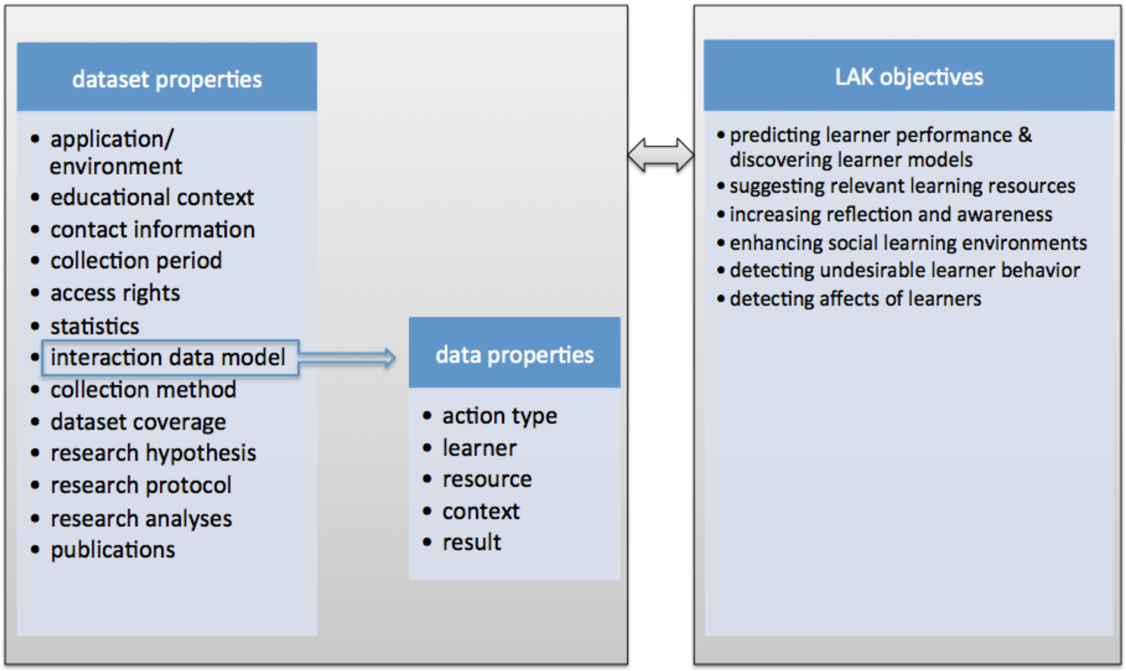


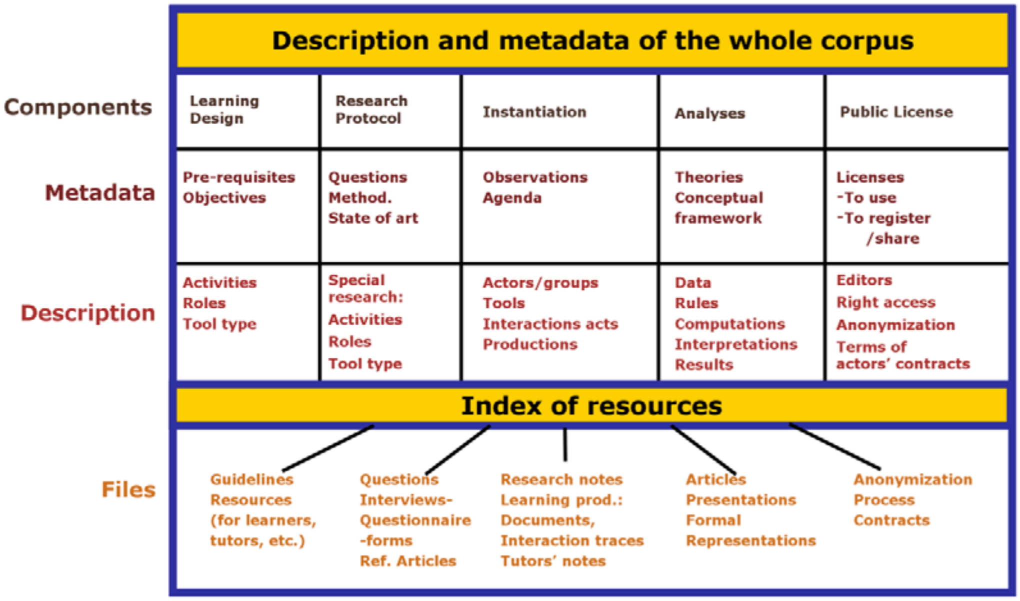
Figure 1. Educational dataset framework

Dataset properties

We recently (Drachsler et al., 2010) presented a specification of datasets that was used for the first dataTEL challenge. Among others, the specification includes information about the application in which the dataset has been collected, the educational setting, contact person, availability (open access or legal protection rules that describe how and when the dataset can be used), dataset collection method, dataset statistics, and pre-processing steps that have been applied to the data.

Related initiatives have also defined formats to package and describe datasets. As illustrated in Figure 2, a Mulce dataset is comprised of the following components:

- The instantiation component includes all interaction data, as well as user information.
- The learning design component describes the educational scenario.
- The research protocol describes the methodology of research with the dataset.
- The license component specifies dataset provider and user rights.
- The analyses component contains research outputs.



Description and metadata of the whole corpus

Components	Learning Design	Research Protocol	Instantiation	Analyses	Public License
Metadata	Pre-requisites Objectives	Questions Method. State of art	Observations Agenda	Theories Conceptual framework	Licenses -To use -To register /share
Description	Activities Roles Tool type	Special research: Activities Roles Tool type	Actors/groups Tools Interactions acts Productions	Data Rules Computations Interpretations Results	Editors Right access Anonymization Terms of actors' contracts

Index of resources

Files	Guidelines Resources (for learners, tutors, etc.)	Questions Interviews-Questionnaire -forms Ref. Articles	Research notes Learning prod.: Documents, Interaction traces Tutors' notes	Articles Presentations Formal Representations	Anonymization Process Contracts

Figure 2. Mulce format (Reffay & Betbeder, 2009)

The PSLC DataShop project defines a specification for describing datasets that are derived from intelligent tutoring systems. The specification includes the project name, principal investigator, curriculum, collection dates, domain, application, description, hypothesis (e.g., "people who are required to use the tutor show less error on quizzes"), school, statistics and knowledge models of interactions. We discuss such interaction models in the next section. In addition, research papers with the datasets are referenced.

As illustrated in Table 1, there are many similarities between the specifications. Explicit information is indicated by "+" signs. This information constitutes explicitly articulated elements of the specifications. Implicit information is indicated by "(+)" signs and represents information that is implied or expressed as part of other elements. For instance, in the dataTEL specification, information about the domain or users can be described as part of the description of the application or environment, but no specific fields are provided for these elements. To date, the Mulce format provides the most comprehensive format for describing datasets. In addition to interaction data, the datasets incorporate a detailed description of the educational scenario in a learning design component and results of various analyses. Therefore, this specification provides the most interesting perspectives for describing educational datasets in a generic way.

Data properties

In addition to a format for describing datasets, there is a need to identify at a more fine-grained level of granularity which data elements are stored. Such information is essential to identify for which research purposes a dataset is useful. As outlined by Romero et al. (2007), the TEL field differs from the e-commerce analytics field in several ways. In e-commerce, the used data are often simple web server access logs or ratings of users on items. In TEL, many researchers use more information about a learner interaction (Pahl & Donnellan, 2002). The user model and the objectives of the systems are also different in both application domains (Drachsler et al., 2009).

Table 1. Overview dataset properties

	dataTEL	Mulce	PSLC DataShop
Application/environment	+	+	+
Educational context	+	+	+
• formal/informal/workplace learning	+	(+)	
• domain/curriculum	(+)	+	+
• school	(+)		+
• learning design		+	
• users	(+)	+	+
Dataset contact information	+	+	+
Collection period	+	+	+
Access rights	+	+	+
Dataset collection method (e.g. logged or user-provided data)	+	+	+
Dataset coverage (overall/instance)	+	+	+
Statistics (number of users, interactions, resources, etc.)	+	+	+
Interaction data model	+	+	+
Raw or cleaned data / preprocessing steps	+	+	+
Research hypothesis		+	+
Research protocol		+	(+)
Research outcomes/analyses		+	
Publications		+	+

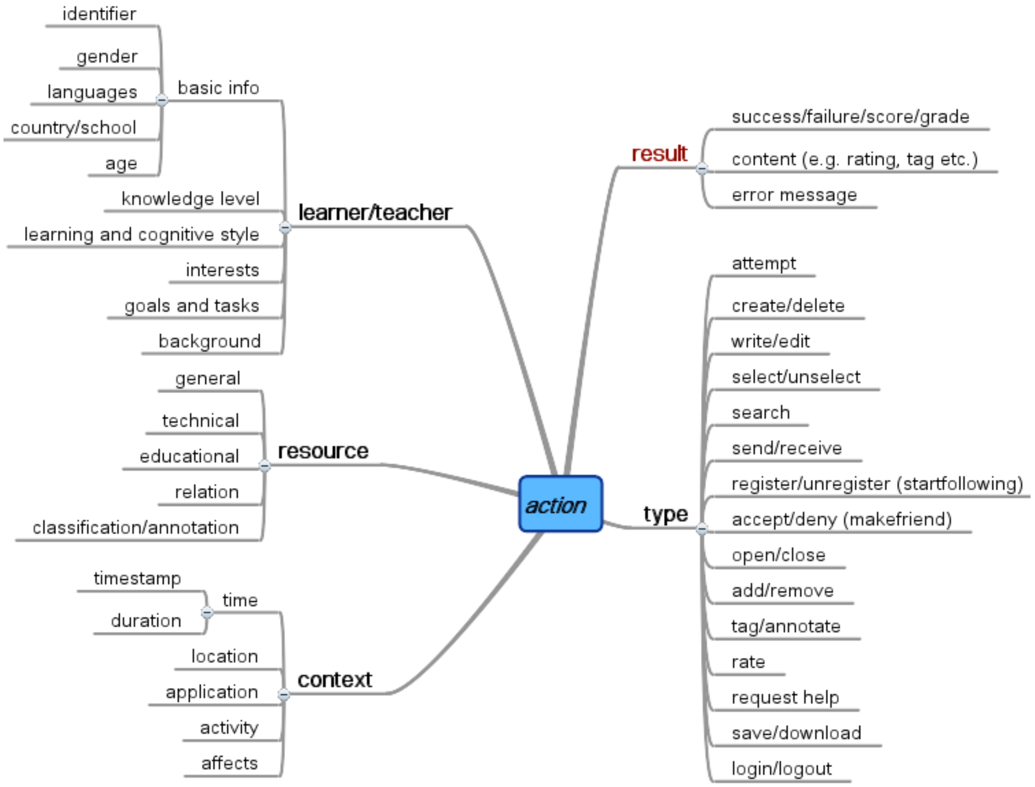


Figure 3. Learner Action Model

A survey of existing TEL interaction data models has been presented in (Butoianu et al., 2010). Such models capture actions of users on resources, such as *open/close*, *select/unselect* or *write* actions, on resources. In addition, the context in which an action occurred can be captured, such as the current application the author is working with or her current task. The Atom activity stream RDF mapping of the LinkedEducation.org initiative provides such a model for actions of users in social networks. Vocabularies for actions, actors and objects involved and related contextual information are defined.

In addition to interaction models, learner models have been elaborated that describe several characteristics of learners. Brusilovsky and Millan (2007) identified the following categories based on an extensive analysis of the literature: *knowledge levels*, *goals and tasks*, *interests*, *background* and *learning and cognitive styles*. In addition, several models, standards and specifications have been elaborated to describe learning resources. The IEEE LOM and Dublin Core metadata standards are prominently used to describe learning resources, including general characteristics, such as title, author and keywords, technical and educational characteristics and relations between learning resources.

We integrated the various data categories and elements in Figure 3. We use this model in the remainder of this article to identify data elements in existing datasets. The model has been developed by synthesizing existing works on interaction data and context variables in the TEL field that were outlined above. It could be further refined by studying relevant theoretical frameworks, like the Activity Theory (Kaptelinin et al., 1995), which could help reorganize the various categories and elements. Future research work in this area is discussed in the last section.

Learning and knowledge analytics objectives

In order to provide guidance on the relevancy of datasets for LAK research, we identify a set of objectives that are relevant for LAK applications. We also outline existing research work in related research communities that, when interconnected, can provide substantial synergies to advance the emerging LAK field.

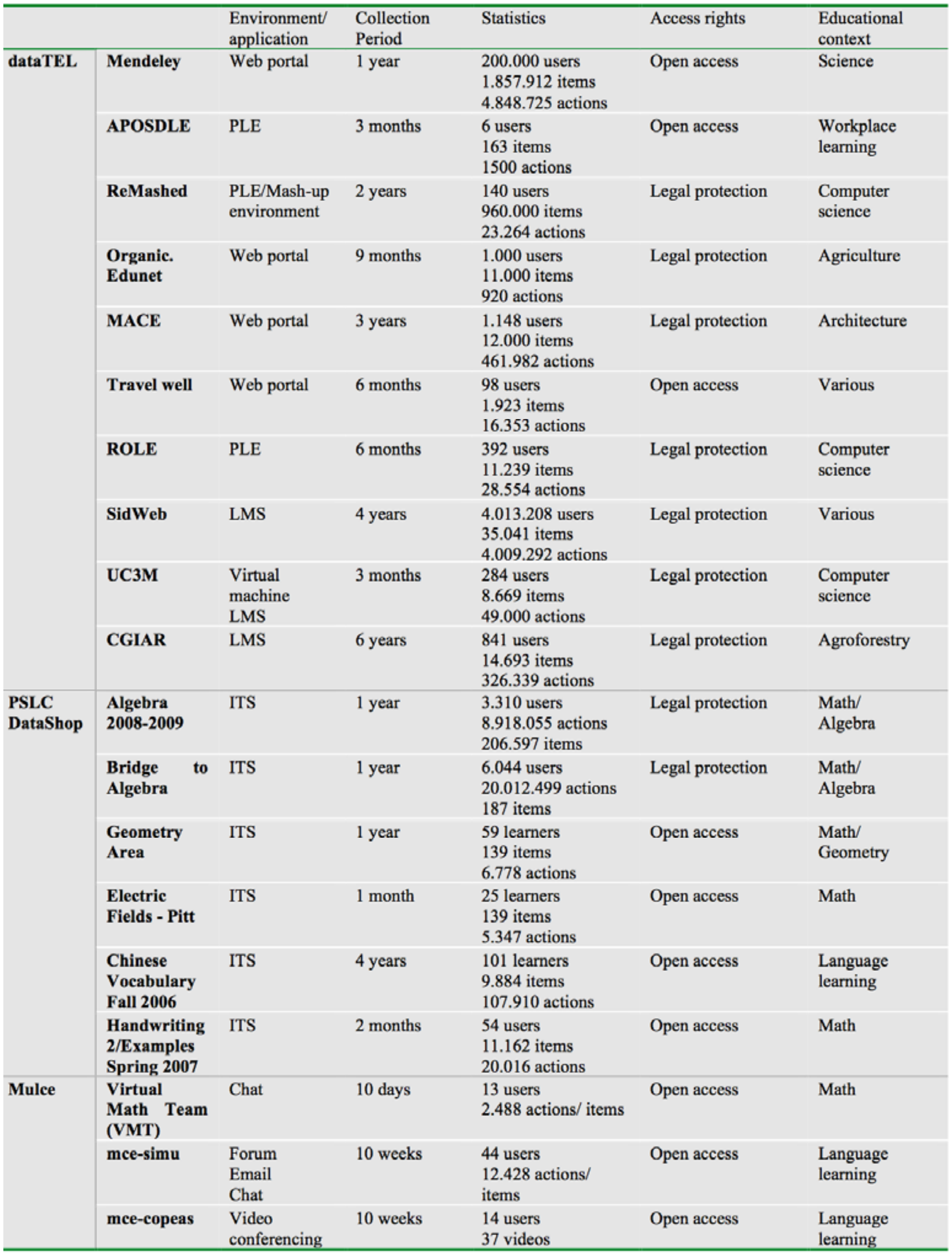
- *Predicting learner performance and modeling learners.* The prediction of learner performance and modeling of learners have been researched extensively by the educational data mining, educational user modeling and educational adaptive hypermedia communities. The objective is to estimate the unknown value of a variable that describes the learner, such as performance, knowledge, scores or learner grades (Romero & Ventura, 2007). Such predictions are for instance used by intelligent tutoring systems to provide advice or hints when a learner is solving a problem. Dynamic learner models are also researched to support adaptation in educational hypermedia systems (Brusilovsky & Millan, 2007).
- *Suggesting relevant learning resources.* Recommender systems for learning have gained increased interest in recent years. A recent survey of TEL recommender systems has been elaborated by Manouselis et al. (2011). These systems typically analyze learner data to suggest relevant learning resources, peer learners or learning paths.
- *Increasing reflection and awareness.* Several researchers are focusing on the analysis and visualization of different learning indicators to foster awareness and reflection about learning processes. These indicators include resource accesses, time spending and knowledge level indicators (Mazza & Milani, 2005).
- *Enhancing social learning environments.* Analysis and visualization of social interactions is researched to make people aware of their social context and to enable them to explore this context (Heer & boyd, 2005). In TEL, this is particularly, but not only, relevant for Computer Supported Collaborative Learning (CSCL) (Stahl, 2009), where the interactions with peer learners are a core aspect of how learning is organized. In CSCL, much research has focused on the analysis of networks of learners, typically with a social network analysis approach (Reffay & Chanier, 2003).
- *Detecting undesirable learner behaviors.* The objective of detecting undesirable learner behavior is to discover those learners who have some type of problem or unusual behavior, such as erroneous actions, misuse, cheating, dropping out or academic failure (Romero & Ventura, 2007).
- *Detecting affects of learners.* Researchers in TEL often refer to the affective states defined by D’Mello et al. (2007). These states are classified as boredom, confusion, frustration, eureka, flow/engagement, versus neutral. Among others, the detection of affects is researched to adjust pedagogical strategies during learning of complex material.

The objectives are highly interrelated. For instance, whereas research on affects and awareness and reflection has traditionally focused on an individual perspective, these objectives are also researched increasingly to enhance social learning environments.

Datasets for Learning and Knowledge Analytics

In this section, we present an analysis of datasets that can be used for a wide variety of LAK research purposes. We analyze the datasets along the dimensions of the dataset framework that we presented in the previous section.

Table 2. Overview dataset properties



		Environment/ application	Collection Period	Statistics	Access rights	Educational context
dataTEL	**Mendeley**	Web portal	1 year	200.000 users 1.857.912 items 4.848.725 actions	Open access	Science
	APOSDLE	PLE	3 months	6 users 163 items 1500 actions	Open access	Workplace learning
	ReMashed	PLE/Mash-up environment	2 years	140 users 960.000 items 23.264 actions	Legal protection	Computer science
	Organic. Edunet	Web portal	9 months	1.000 users 11.000 items 920 actions	Legal protection	Agriculture
	MACE	Web portal	3 years	1.148 users 12.000 items 461.982 actions	Legal protection	Architecture
	Travel well	Web portal	6 months	98 users 1.923 items 16.353 actions	Open access	Various
	ROLE	PLE	6 months	392 users 11.239 items 28.554 actions	Legal protection	Computer science
	SidWeb	LMS	4 years	4.013.208 users 35.041 items 4.009.292 actions	Legal protection	Various
	UC3M	Virtual machine LMS	3 months	284 users 8.669 items 49.000 actions	Legal protection	Computer science
	CGIAR	LMS	6 years	841 users 14.693 items 326.339 actions	Legal protection	Agroforestry
PSLC DataShop	**Algebra 2008-2009**	ITS	1 year	3.310 users 8.918.055 actions 206.597 items	Legal protection	Math/ Algebra
	Bridge to Algebra	ITS	1 year	6.044 users 20.012.499 actions 187 items	Legal protection	Math/ Algebra
	Geometry Area	ITS	1 year	59 learners 139 items 6.778 actions	Open access	Math/ Geometry
	Electric Fields - Pitt	ITS	1 month	25 learners 139 items 5.347 actions	Open access	Math
	Chinese Vocabulary Fall 2006	ITS	4 years	101 learners 9.884 items 107.910 actions	Open access	Language learning
	Handwriting 2/Examples Spring 2007	ITS	2 months	54 users 11.162 items 20.016 actions	Open access	Math
Mulce	**Virtual Math Team (VMT)**	Chat	10 days	13 users 2.488 actions/ items	Open access	Math
	mce-simu	Forum Email Chat	10 weeks	44 users 12.428 actions/ items	Open access	Language learning
	mce-copeas	Video conferencing	10 weeks	14 users 37 videos	Open access	Language learning

Dataset properties

Table 2 provides an overview of characteristics of available educational datasets, including the application from which data were collected, collection period, statistics and educational context or domain. The full description of the datasets is available on the portals that provide access to these datasets, including the dataTEL (http://www.teleurope.eu/pg/pages/view/50630/), DataShop (https://pslcdatashop.web.cmu.edu/) and Mulce (http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/) portals.

Several dataTEL datasets have been collected from *learning management systems (LMS)*. The UC3M dataset also collects data from a *virtual machine* that was used in a C programming course. The particularity of this dataset is that it records actions from several tools learners are using. The approach enables to collect a more comprehensive overview of learner activities, such as a learner searching for additional resources on the web. Such an approach is also researched under the prism of *personal learning environments (PLEs)*, where data is tracked from learning environments that assemble relevant tools for course activities. Many other dataTEL datasets were collected from *web portals* that provide access to large collections of learning resources.

Several other datasets are collected from *intelligent tutoring systems (ITS)* – including a large number of datasets from the PSLC DataShop initiative. We include the “Algebra 2008-2009” and “Bridge to Algebra” datasets that were used for the KDD 2010 Cup on educational data mining (https://pslcdatashop.web.cmu.edu/KDDCup/) in this analysis. In addition, the recommended datasets of the DataShop are analyzed. At the time of writing, 64 datasets are publicly available. Finally, many of the Mulce datasets contain data that were captured from *forums, chat* and *email* conversations between learners in collaborative learning settings.

The collection period varies from 10 days to 6 years. Several of the Mulce datasets capture data of group work during a specific learning activity. Datasets derived from learning management systems and web portals often capture data during a longer period of time, ranging from a couple of months to several years. Although a few datasets that are available capture data of a large number of users, many other datasets are more limited in size. Some datasets collect data of 1000 to 7000 users. Several other datasets capture data of a few learners only. These datasets are in some cases only a sample that the organization made available or in other cases datasets of a small number of collaborating users, such as the VMT and mce-copeas datasets.

Several datasets are openly accessible. For other datasets, legal protection rules apply. We obtained these datasets by sending a statement of our intended research purposes to the organization and then signed an agreement on the use of these data. All datasets contain data that is anonymized, so that it can no longer be linked to an individual.

Data properties

Table 3 presents a more detailed overview of the data elements that are included in the datasets. The datasets contain a diverse set of *actions* of users. These actions include attempts of learners on quizzes, search actions, selection, annotation, rating, creation or editing of resources. PSLC datasets derived from intelligent tutoring systems all include attempt actions on activities provided by the tutor. In some datasets, help requests are stored. The input provided by learners is sometimes further specified into select, write or create actions. The Mulce datasets capture social interactions – in most cases these constitute send and receive actions.

Explicit information about *learners* (or *teachers*) is stored in only a few datasets. The data is in most cases anonymized and little additional information about learners or teachers is stored. Some dataTEL datasets contain information about the language, interests, knowledge level or country of the user. Some DataShop datasets describe the gender and knowledge level of the learner, including her past grades. The mce-copeas dataset divides learners in three groups according to their knowledge level (beginner, medium, expert). Information about country, age, language and gender is often provided in Mulce datasets.

Information about *resources* is available in more datasets. The information ranges from an identifier of the resource to detailed descriptions that include educational characteristics such as duration, minimum age, maximum age and resource type, technical characteristics and annotations such as tags and comments. Such metadata are often provided in dataTEL datasets that were captured from learning repository portals. In the DataShop datasets, educational information such as average duration and required skills are sometimes provided. In addition, compositional relations are provided that define a hierarchy of units and sections. Social relations between learners collaborating are stored in the Mulce and some dataTEL datasets.

Additional *context* information is also stored. Several datasets provide timestamp information. The duration of an action is stored explicitly as a time interval in the DataShop and some LMS datasets. Such information is valuable to calculate the difference of estimated durations, described in resource metadata, and the time the learner needed in practice to complete an activity. Other contextual information is not often available. In datasets that contain data of multiple tools and services, information about the application from which an action was triggered is included.

Table 3. Overview data properties

		dataTEL										PSLC dataShop						Mulce		
		Mendeley	APOSDLE	ReMashed	Organic.Edunet	MACE	Travel well	ROLE	SidWeb	UC3M	CGIAR	Algebra 2008-2009	Bridge to Algebra	Geometry area	Electric Fields - Pitt	Chinese Vocabulary	Handwriting2/Example	VMT	mce-simu	mce-copeas
action type	attempt		+						+	+	+	+	+	+	+	+	+			
	create/delete		+			+		+	+	+	+				+					
	write/edit		+							+	+	+			+					+
	select/unselect	+	+			+		+	+	+	+	+			+					+
	search		+			+		+		+	+									
	send/receive		+					+	+		+							+	+	+
	register/unregister		+			+					+									
	open/close		+					+		+	+									
	add/remove					+					+									+
	tag/annotate		+	+	+	+	+				+									
	rate			+	+	+	+													
	request help										+	+	+		+					
	save/download	+	+			+	+				+									
	login/logout			+		+					+							+		
learner /teacher	id	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	country/school						+				+						+		+	+
	gender																+		+	+
	age																		+	+
	languages						+				+								+	
	knowledge level			+											+		+			+
	interest			+			+													
	goals/tasks																			
	background																			
resource	general	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	technical		+			+	+		+											
	educational					+	+					+	+	+	+	+	+			
	relation		+			+			+		+	+	+	+	+	+		+	+	+
	classification/annotation					+	+													
context	timestamp		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
	duration								+		+	+	+	+	+	+	+	+		+
	application							+		+									+	
result	error message									+					+					
	success/failure											+	+	+	+	+	+			
	score/grade										+						+			
	content (rating value, tag, etc.)		+	+	+	+	+	+	+	+	+						+	+	+	+

Finally, the *result* of actions, such as correct or incorrect attempts, rating values or error messages, is stored. In addition, some datasets contain the grade a learner obtained for an activity or course. We elaborate in the next section how such data can be used for LAK research.

Usefulness of available datasets for LAK related research objectives

Prediction of learner performance and discovering learner models

Several datasets are available that can support research on prediction of learner performance and discovery of learner models. Among others, such predictions are researched to provide advice when a learner is solving a problem (Romero et al., 2007). Datasets from intelligent tutoring systems that capture attempts of learners provide a rich source of data to estimate the knowledge level of a learner. Some datasets derived from LMSs contain data on the number of attempts and total time spent on assignments, forums and quizzes. Romero et al. (2008) compared different data mining techniques to classify learners based on such LMS data and the final grade obtained for courses. Datasets that have been captured from PLEs offer interesting perspectives to elaborate such research in open learning environments. In addition, many datasets are suitable to identify interests of users based on resource accesses.

Several researchers have already experimented with the datasets outlined above to predict learner attributes. The “Algebra 2008-2009” and “Bridge to Algebra” datasets were used in the KDD Challenge 2010. Participants were asked to learn a model from past learner behavior and to predict their future performance. The winners of this competition combined several educational data mining techniques. Cen et al. (2007) performed a learning curve analysis with the “Geometry Area” dataset. They noticed that while learners were required to over-practice some easy target skills, they under-practiced harder skills. Based on this observation, they created a new version of the tutor by resetting parameters that determine how often skills are practiced. References to other studies with these datasets are available on the DataShop portal.

Suggesting learning resources

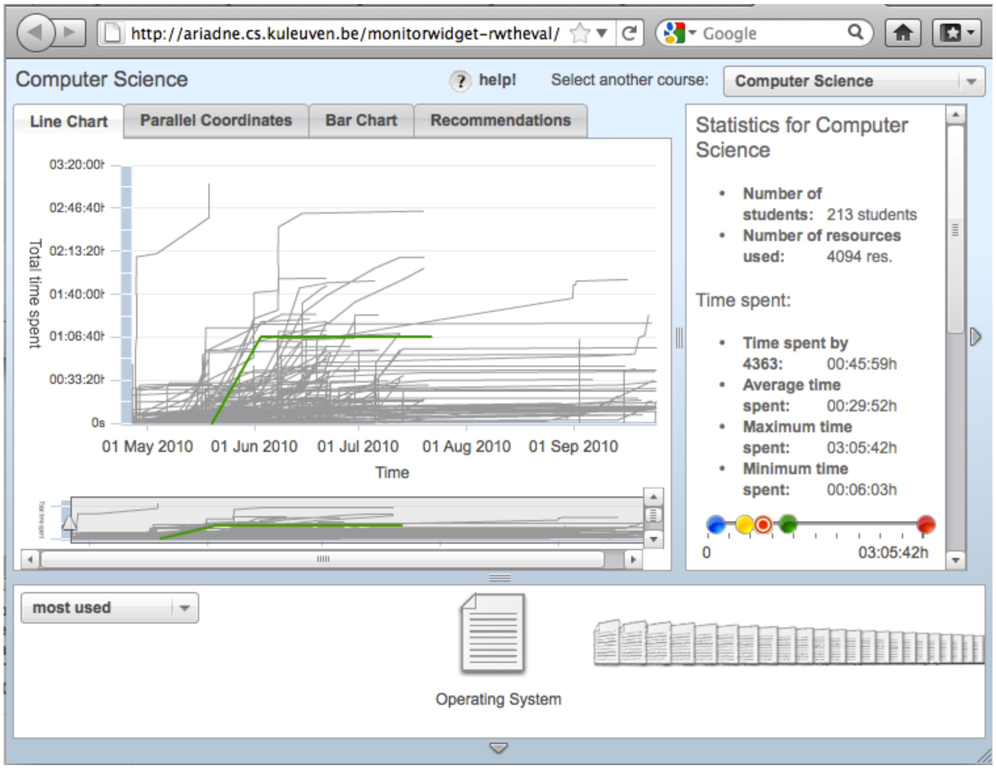
Several dataTEL datasets contain explicit relevance indicators in the form of ratings that are relevant for research on recommendation algorithms for learning. In addition, implicit relevance indicators, such as downloads, search terms and annotations, are available that can be used for such research. If time interval data is available, the data might be suitable to extract reading times in order to determine the relevancy of a resource. In addition, such datasets are useful to analyze information about sequences of resources as a basis to suggest learning paths.

Manouselis et al. (2010b) used the Travel well dataset to evaluate recommendation algorithms for learning. Similar experiments have been reported in (Verbert et al., 2011). In this study, the Mendeley and MACE datasets were also used. Although still preliminary, some conclusions were drawn about successful parameterization of collaborative filtering algorithms for learning. Outcomes suggest that the use of implicit relevance indicators, such as downloads, tags and read actions, are useful to suggest learning resources.

Increasing reflection and awareness

Several datasets are useful for analysis and visualization of different learning indicators to foster awareness and reflection about learning processes. In addition to indicators about the knowledge level of learners, several datasets contain indicators of the time learners spend on learning activities – such as the PSLC DataShop datasets.

Other datasets contain timestamp information that can be used to derive indicators of the time users were active. dataTEL datasets were for instance used to obtain such indicators as a basis to support awareness for teachers (Govaerts et al. 2010). A visualization of these indicators applied to the ROLE dataset is illustrated in Figure 4. Evaluation results indicate that the perceived usefulness for teachers is high. The MACE dataset has been used for research on reflection and awareness of resource accesses. The Zeitgeist application (Shmidz et al., 2009) gives users insight into which learning resources they accessed, how they found them and which topics have been of interest to them (see Figure 5).



Computer Science

Select another course: Computer Science

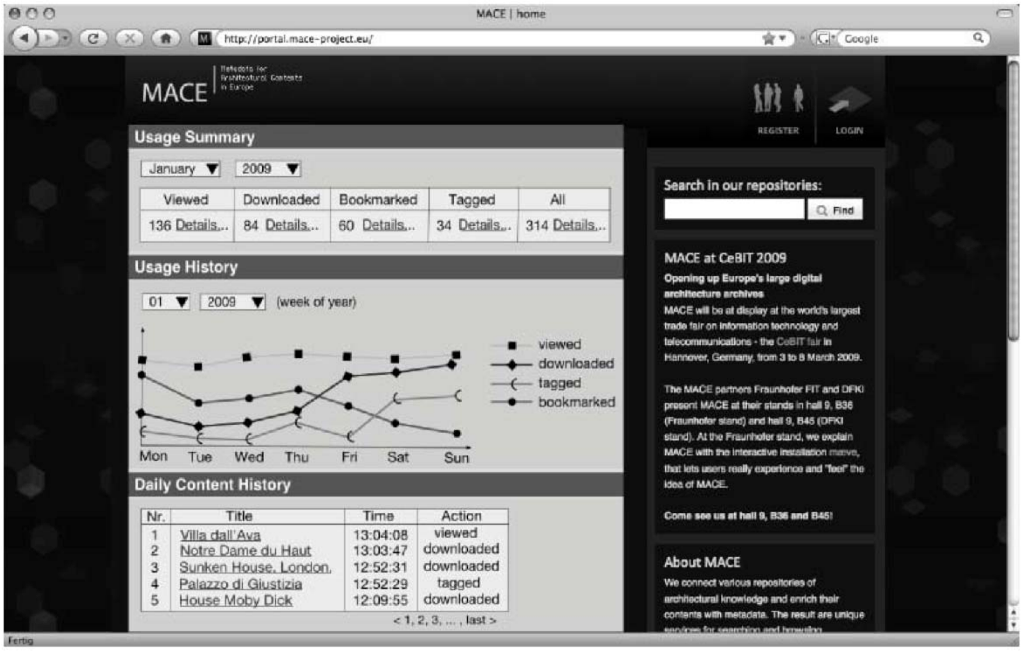
Statistics for Computer Science

- Number of students: 213 students
- Number of resources used: 4094 res.

Time spent:

- Time spent by 4363: 00:45:59h
- Average time spent: 00:29:52h
- Maximum time spent: 03:05:42h
- Minimum time spent: 00:06:03h

Figure 4. Visualization of time indicators (Govaerts et al., 2010)



Usage Summary

January 2009

Viewed	Downloaded	Bookmarked	Tagged	All
136 Details...	84 Details...	60 Details...	34 Details...	314 Details...

Usage History

01 2009 (week of year)

Daily Content History

Nr.	Title	Time	Action
1	Villa dall'Ava	13:04:08	viewed
2	Notre Dame du Haut	13:03:47	downloaded
3	Sunken House, London.	12:52:31	downloaded
4	Palazzo di Giustizia	12:52:29	tagged
5	House Moby Dick	12:09:55	downloaded

< 1, 2, 3, ... , last >

Search in our repositories:

MACE at CeBIT 2009

Opening up Europe's large digital architecture archives
MACE will be at display at the world's largest trade fair on information technology and telecommunications - the CeBIT fair in Hannover, Germany, from 3 to 8 March 2009.

The MACE partners Fraunhofer FIT and DFKI present MACE at their stands in hall 9, B36 (Fraunhofer stand) and hall 9, B45 (DFKI stand). At the Fraunhofer stand, we explain MACE with the interactive installation maeve, that lets users really experience and "feel" the idea of MACE.

Come see us at hall 9, B36 and B45!

About MACE

We connect various repositories of architectural knowledge and enrich their contents with metadata. The result are unique services for searching and browsing

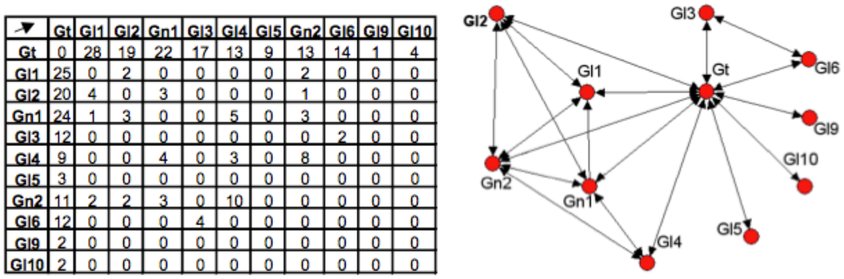
Figure 5. MACE Zeitgeist (Shmidz et al., 2009)

Enhancing social learning environments

Several Mulce datasets are useful for research on collaborative learning. The datasets have been captured from chat tools, forums or email clients. Such data can be analyzed to predict and advice on learning in group work. Datasets that have been captured from LMSs often capture messages within course forums. Some of the PSLC DataShop datasets capture collaborative activities with intelligent tutoring systems, including the “Electric Fields – Pitt” dataset.

Several datasets have already been used to support research on enhancing social learning environments. Research with the “Electric Fields – Pitt” dataset suggests that asking learners to solve problems collaboratively with an intelligent tutoring system is a productive way to enhance learning from an ITS (Hausmann et al., 2008).

Several Mulce datasets have been used for research on collaborative learning (Stahl, 2009). Among others, the datasets have been used to understand mathematical ideas and reasoning in chat by learners, interaction mechanisms used by online groups to sustain knowledge building over time and the measurement of cohesion in collaborative distance learning. Evaluation studies showed that such analysis, when embodied in visualization tools (see Figure 6), can efficiently assist the teacher in following the group collaboration (Reffay & Chanier, 2003). These analyses were used to highlight isolated people, active sub-groups and various roles of the members in group communication. The mce-copeas dataset has been used to research the influence of synchronous communication during online collaborative writing activities (Ciekanski & Chanier, 2008). Several other studies are documented on the Mulce portal.



↘	Gt	Gl1	Gl2	Gn1	Gl3	Gl4	Gl5	Gn2	Gl6	Gl9	Gl10
Gt	0	28	19	22	17	13	9	13	14	1	4
Gl1	25	0	2	0	0	0	0	2	0	0	0
Gl2	20	4	0	3	0	0	0	1	0	0	0
Gn1	24	1	3	0	0	5	0	3	0	0	0
Gl3	12	0	0	0	0	0	0	0	2	0	0
Gl4	9	0	0	4	0	3	0	8	0	0	0
Gl5	3	0	0	0	0	0	0	0	0	0	0
Gn2	11	2	2	3	0	10	0	0	0	0	0
Gl6	12	0	0	0	4	0	0	0	0	0	0
Gl9	2	0	0	0	0	0	0	0	0	0	0
Gl10	2	0	0	0	0	0	0	0	0	0	0

Figure 6. Matrix and graphical representation of e-mail exchange (Reffay & Chanier, 2003)

Detecting undesirable learner behaviors

Datasets derived from PLEs provide a rich source of data to detect unusual behavior, as these datasets record actions of learners with several tools they were using during the classes. Data from LMSs can be used to detect potential dropouts when learners are no longer active. ITS datasets are also suitable for research on unusual behavior. Baker et al. (2006) found that learners who were “gaming the system” (i.e., fast and repeated requests for help to avoid thinking) had the largest correlation with poor learning outcomes.

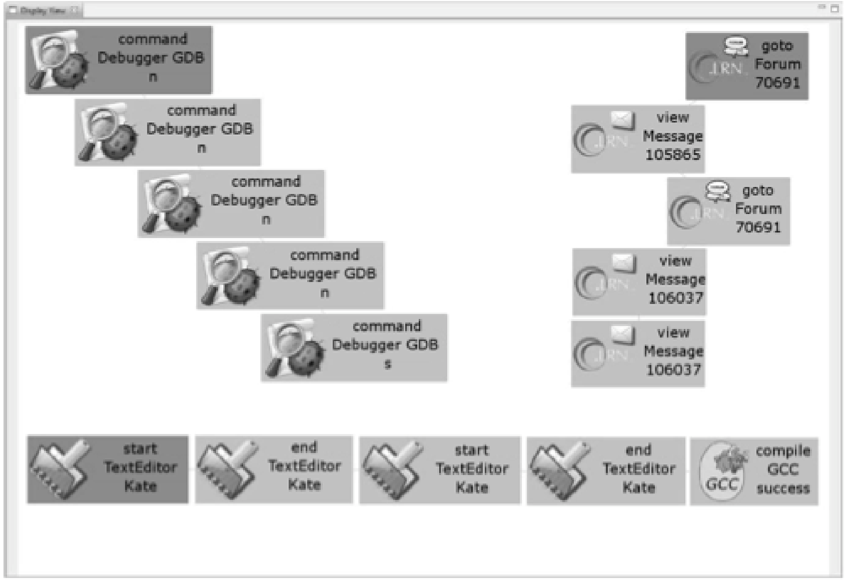


Figure 6. Pattern visualization of UC3M dataset (Scheffel et al., 2011)

Scheffel et al. (2011) used the UC3M dataset to identify key actions from observed learning behavior. The authors employed data mining techniques to extract frequent patterns of actions. These patterns were visualized to support teaching activities. For instance, the pattern illustrated in Figure 7 points to development flows in which for each compilation students opened a file and closed it again before compiling. According to the teaching staff, such actions translate into a significant increase in development time and should be corrected.

Detecting affects of learners

Some datasets are suitable for research on the detection of affects and motivational aspects. For instance, PSLC DataShop datasets can be used to extract motivational aspects by comparing the time a learner spends on a learning activity in an ITS with the expected or average time of other learners. The use of emoticons and affective words is researched in social interaction datasets (Reffay et al., 2011). Prominent research in building a user model of affect from real data has been conducted by Conati and Maclaren (2005).

Ongoing research with the UC3M dataset is focused on the detection of affects of learners, such as frustration and (dis-)engagement. Based on an analysis of sequences of actions, such as a sequence of error messages of a debugger and successful compilations, information is deduced about potential engagement or frustration.

Conclusion and future challenges

In this article, we have presented an overview of datasets that can be used for exploratory research on LAK. Several datasets have been identified and analyzed along the dimensions of our educational dataset framework. Our analysis indicates that an initial collection of interesting datasets is already available. These datasets have been used in a successful way by several researchers, for diverse research purposes that are relevant for LAK analytics. The analysis provides researchers with an overview of available datasets, as well as their properties to help decide about their appropriateness for their experiments. In addition, the analysis sheds some light on what data to track for LAK research. The datasets were collected by our dataTEL and related initiatives, with the common objective to make educational datasets available to researchers.

Nevertheless, our endeavors to collect and share datasets for research remain quite challenging. A first challenge is related to privacy rights and licensing of educational data. Although an enormous amount of data has been captured from learning environments, it is a difficult process to make these data available for research purposes. The issue of usage rights/licensing needs to be solved from two perspectives. From a user perspective, learners need to be informed and grant permission to collect their data and make it available for research purposes. Also the organization or provider of these data needs to agree with collecting and sharing these data. For instance, researchers have in some cases collected datasets by crawling data from websites and then found out that they were not allowed to do so.

The collection of the UC3M dataset is a good example of how data can be made available for research purposes. In a first stage, learners were informed about which data were collected during their course activities and signed an agreement that the data could be used for research purposes, as is required by the Spanish law on data protection and privacy. In a second stage, researchers signed an agreement on the use of the dataset for research purposes. In order to facilitate the collection of educational datasets, it is important to share such practices and to contribute to the documentation of legal data protection and privacy laws. In addition, providing guidelines for anonymization of data and creating incentives for researchers and organizations to share their datasets is important. The first dataTEL challenge requested submissions of dataset fact sheets in a pre-defined format so that required properties could be well documented. The approach is similar to work of the DataVerse Network (King, 2007).

A second challenge is related to the heterogeneity of educational datasets. The lack of a standard representation for interaction data within datasets prevents the sharing and reuse of data across systems. In addition, when a custom data format is not well documented, it may be difficult to assess the meaning and usefulness of data elements that are stored. To address this issue, the development of a standard representation for learner interaction data will be taken up by a working group of the CEN Workshop on Learning Technologies (WS/LT).

This challenge sets the context of a third challenge, the identification of relevant data about learners and other entities involved for LAK research. Whereas we were able to identify some data elements that can be used as input for the CEN WS/LT working group, additional research is required to identify a broader set of elements that are useful for LAK research. In addition, the model could be further refined by studying relevant theoretical frameworks, like the Activity Theory (Kaptelinin et al., 1995), which could help us reorganize the various categories and elements according to well-established frameworks.

A fourth challenge is the development of data sensors to collect data. Data that is tracked within learning management systems provides a good basis for exploratory research on learning and knowledge analytics. However, learners often use a wide variety of tools and services in addition to a traditional LMS. Ongoing work within the ROLE project (http://www.role-project.eu) is focused on collecting data from PLEs that aggregate several tools and services. The approach is inspired by wakoopa (http://social.wakoopa.com) and rescuetime (http://www.rescuetime.com) that install tracker tools on the machine of a user and automatically record all activities.

In addition, data sharing and reuse in the educational field needs further research to explore whether the context and scope of the dataset collection can significantly affect its potential reuse. This is a requirement to keep in mind, as relevant discussions on productive multivocality in CSCL have indicated that there is a possibility that data collected using one theoretical framing may be unsuitable for analysis under another framing. The issue also indicates that clear descriptions of datasets along various dimensions describing such properties, as provided by the Mulce and DataShop specifications, are important for exchanging and possibly reusing datasets that have been collected in various settings.

Finally, the interconnection of several efforts in the area of educational dataset collection is key to advance this work. dataTEL has recently been accepted as an EATEL Special Interest Group (SIG). The dataTEL SIG aims to bring together existing efforts in the area of dataset collection and sharing. The SIG will organize yearly workshops and three monthly virtual meetings. With the organization of these events, we hope to enable collection and sharing of datasets on a much larger scale than available today.

Acknowledgements

Part of this work was supported by the dataTEL Theme Team of the FP7 STELLAR Network of Excellence (no. 231913). We thank all members involved, and in particular Martin Wolpers, Miguel-Angel Sicilia, Stefanie Lindstaedt and Riina Vuorikari, for their contributions to the organization of dataTEL events. Katrien Verbert is a Postdoctoral Fellow of the Research Foundation-Flanders (FWO). The work of Nikos Manouselis has been funded with support by the EU project VOA3R-250525 of the CIP PSP Programme (http://voa3r.eu).

References

Baker, R. S. J., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., … Beck, J. (2006). Adapting to when students game an intelligent tutoring system. In M. Ikeda et al. (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 392-401). Berlin: Springer.

Brusilovsky, P., & Millan, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovsky et al. (Eds.), *The adaptive web* (pp. 3–53). Berlin, Germany: Springer.

Butoianu, V., Vidal, P., Verbert, K., Duval, E., & Broisin, J. (2010). User context and personalized learning: A federation of contextualized attention metadata. *Journal of Universal Computer Science, 16*(16), 2252-2271.

Cen, H., Koedinger, K., & Junker, B. (2007). Is over practice necessary? – Improving learning efficiency with the cognitive tutor through educational data mining. In R. Luckin & K. Koedinger (Eds.), *Proceedings of the 13th International Conference on AIED* (pp. 511-518). Los Angeles: IOS Press.

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications, 23*(3), 187-200.

Ciekanski, M., & Chanier, T. (2008). Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. *ReCALL, 20*(2), 162-182.

Conati, C., & Maclaren, H. (2005). Data-driven refinement of a probabilistic model of user affect. In L. Ardissono, P. Brna, & A. Mitrovic (Eds.), *User Modeling 2005* (pp. 40-49). Berlin: Springer.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Towards an affect-sensitive autotutor. *IEEE Intelligent Systems, 22*(4), 53-61.

Drachsler, H., Hummel, H. G. K., & Koper, R. (2009). Identifying the goal, user model and conditions of recommender systems for formal and informal learning. *Journal of Digital Information, 10*(2), 4-24.

Drachsler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., … Wolpers, M. (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science, 1*(2), 2849-2858.

Duval, E. (2011). Attention please!: Learning analytics for visualization and recommendation. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge,* (pp. 9-17). New York, NY: ACM.

Elias, T. (2011). *Learning Analytics: Definitions, processes and potential.* Retrieved February 9, 2012, from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Govaerts, S., Verbert, K., Klerkx, J., & Duval, E. (2010). Visualizing activities for self-reflection and awareness. *Lecture Notes in Computer Science, 6483,* 91-100.

Hausmann, R. G. M., van de Sande, B. & VanLehn, K. (2008). Trialog: How peer collaboration helps remediate errors in an ITS**.** In D. Wilson & H. C. Lane (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference* (pp. 415-420). Menlo Park, CA: AAAI Press.

Heer, J., & boyd, d. (2005). Vizster: Visualizing online social networks. In J. T. Stasko & M. O. Ward (Eds), *Proceedings of the 2005 IEEE Symposium on Information Visualization* (p. 5). Washington, DC: IEEE.

Kaptelinin, V., Kuutti, K., & Bannon, L. J. (1995). Activity theory: Basic concepts and applications. In B. Blumenthal et al. (Eds.), *Selected papers from the 5th International Conference on Human-Computer Interaction* (pp. 89-201). London, UK: Springer-Verlag.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods Research*, *36*(2), 173-199.

MacWhinney, B. (1996). The CHILDES System. *American Journal of Speech-Language Pathology, 5,* 5-14.

MacWhinney, B. (2007). *The TalkBank project.* Retrieved from Carnegie Mellon University, Department of Psychology website: http://repository.cmu.edu/psychology/174

Manouselis, N., Drachsler, H., Verbert, K., & Santos, O. C. (2010a). RecSysTEL preface 2010. *Procedia Computer Science, 1*(2), 2773–2774.

Manouselis, N., Vuorikari, R., & Van Assche, F. (2010b). Collaborative recommendation of e-learning resources: An experimental investigation. *Journal of Computer Assisted Learning, 26(4),* 227–242.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., & Koper, R. (2011). Recommender systems in technology enhanced learning. In L. Rokach et al. (Eds.), *Recommender systems handbook: A complete guide for research scientists & practitioners* (pp.387-415). Berlin, Germany: Springer.

Mazza, R., & Milani, C. (2005, July). *Exploring usage analysis in learning systems: Gaining insights from visualisations.* Paper presented at the Workshop on Usage Analysis in Learning Systems, Amsterdam, The Netherlands.

Pahl, C., & Donnellan, C. (2002). Data mining technology for the evaluation of web-based teaching and learning systems. In M. Driscoll & T. Reeves, (Eds.), *Proceedings of Conference on E-Learning in Business* (pp. 747-752). Chesapeake, VA: AACE.

Reffay, C., & Chanier, T. (2003, June). *How social network analysis can help to measure cohesion in collaborative distance learning.* Paper presented at the Conference on Computer Supported Collaborative Learning, Bergen, Norway.

Reffay, C., Teplovs, C., & Blondel, F. -M. (2011). Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. In H. Spada et al. (Eds.), *Proceedings of CSCL 2011*(pp. 846-850). Hong Kong, China: International Society of the Learning Sciences.

Reffay, C., & Betbeder, M. -L. (2009). Sharing corpora and tools to improve interaction analysis. In U. Cress, V. Dimitrova, & M. Specht (Eds.), *Proceedings of EC-TEL '09, LNCS*, volume (pp. 196-210). Berlin, Germany: Springer-Verlag.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Romero, C., Ventura, S. Espejo, P.G., & Hervs, C. (2008). Data mining algorithms to classify students. Retrieved from 1st International Conference on Educational Data Mining website: http://www.educationaldatamining.org/EDM2008/index.php?page=proceedings

Scheffel, M., Niemann, K., Pardo, A., Leony, D., Friedrich, M., Schmidt, K., … Kloos, C.D. (2011). Usage pattern recognition in student activities. In U. Cress et al. (Eds), *Proceedings of Fourth European Conference on Technology Enhanced Learning,* (pp.341-355). Berlin: Springer-Verlag

Schmitz, H.-C., Scheffel, M., Friedrich, M., Jahn, M., Niemann, K. & Wolpers, M. (2009). CAMera for PLE. *Lecture Notes in Computer Science, 5794*, 507–520.

Siemens, G. (2010). What are Learning Analytics? Retrieved August 9, 2011, from http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

Siemens, G., & Gasevic, D. (2011). *Proceedings of the 1st conference on Learning Analytics and Knowledge 2011* (pp. 1-185). New York, NY: ACM.

Stahl, G. (2009). Studying virtual math teams. New York, NY: Springer

Stamper, J. C., Koedinger, K. R., Baker, R. S. J. D., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A data analysis service for the learning science community. In V. Aleven et al. (Eds.), *Proceedings of Intelligent Tutoring Systems* (pp. 455-456). Berlin: Springer.

Treloar, A., & Wilkinson, R. (2008). Access to data for eResearch: Designing the Australian national data service discovery services. *International Journal of Digital Curation, 3*(2)*,* 151-158.

Verbert, K., Drachsler, H., Manouselis, N., Wolpers, M., Vuorikari, R., & Duval E. (2011). Dataset-driven research for improving recommender systems for learning. In G. Siemens & D. Gasevic (Eds.), *Proceedings of the 1st Learning Analytics & Knowledge Conference* (pp. 44-53). New York, NY: ACM.