Preface

Hendrik Drachsler*

Centre for Learning Sciences and Technologies (CELSTEC),
Open University in the Netherlands,
PO Box 2960, 6401 DL Heerlen, The Netherlands
Email: hendrik.drachsler@ou.nl
*Corresponding author

Katrien Verbert

Departement Computerwetenschappen,
KU Leuven,
Celestijnenlaan 200A, B-3001 Leuven, Belgium
Email: Katrien.verbert@cs.kuleuven.be

Nikos Manouselis

Agro-Know Technologies,
17 Grammou Str., Vrilissia Attikis,
15235 Athens, Greece
Email: nikosm@ieee.org

Riina Vuorikari

European Schoolnet,
Rue de Trèves, 61,
1040 Brussels, Belgium
Email: Riina.Vuorikari@eun.org

Martin Wolpers

Fraunhofer-Institut für Angewandte Informationstechnik FIT,
Schloss Birlinghoven,
53754 Sankt Augustin, Germany
Email: martin.wolpers@fit.fraunhofer.de

Stefanie Lindstaedt

Know-Center,
Inffeldgasse 21a, 8010 Graz, Austria
Email: slind@know-center.at

Biographical notes: Hendrik Drachsler is Assistant Professor at the CELSTEC institute of the Open University in the Netherlands. He is focusing on the personalisation of learning with information retrieval technologies and especially recommender systems. Therefore, he is interested in research on educational datasets, linked data, data mashups, data visualisations and learning analytics. He applied these technologies in different international research projects. He organises an annual workshop series on recommender systems and data-driven research for TEL (SIRTEL08-09, RecSysTEL2010, 2012, dataTEL), is co-author of the book on 'Recommender Systems for Learning' and has published numerous papers at international conferences and in well-renown journals.

Katrien Verbert is a post-doctoral researcher of the Research Foundation – Flanders (FWO) at the HCI research unit of KU Leuven, Belgium. Her research interests include content models, content reusability, context-aware recommendation and personalisation, and applications thereof in TEL and science information systems. In that respect, she is currently involved with the RAMLET IEEE LTSC standardisation project and the EU FP7 project ROLE that is focusing on contextual recommendation in learning environments. She co-organised the workshops on Context-Aware Recommendation for Learning, and the RecSysTEL 2010, 2012 workshop and is co-author of the book on 'Recommender Systems for Learning'.

Nikos Manouselis is the R&D Director of Agro-Know Technologies, a research-oriented SME focusing on knowledge-intensive technology innovation for agriculture and rural development. He is also giving lectures on topics related to Web Science and Web Technology at the University of Alcala, Spain. He is experienced in the design, implementation, and coordination of initiatives that are deploying recommendation services on federations of learning repositories for specific user communities. He organises an annual workshop series on recommender systems for TEL (SIRTEL08-09, RecSysTEL 2010, 2012) and is the main author of the book on 'Recommender Systems for Learning'.

Riina Vuorikari works in the field of TEL since 2000. Her areas of interest range from pedagogical innovation to school collaboration and to systems that support social learning opportunities. She currently holds a part-time expert position at European Schoolnet and manages a LLP project (Tellnet) that applies social network analysis, information visualisation and learning analytics to a community of more than 150,000 teachers. She also runs her own consulting business and co-chaired the Social Information Retrieval for Technology Enhanced Learning workshops (SIRTEL) in 2007 to 2009, and coedited a special issue on the topic.

Martin Wolpers holds a PhD in Electrical Engineering and Information Technology from the Leibnitz University Hannover. He is head of the research department 'Context and Attention in Personalised Learning Environments' at the Fraunhofer Institute for Applied Information Technology FIT in Sankt Augustin, Germany. He also holds the title 'visiting professor' at the Katholieke Universiteit Leuven, Belgium. He is strongly involved in a number of highly successful European Projects. He is vice-president of the European Association of TEL (EATEL) and president of the MACE Association. His research interests focuses on the improvement of TEL by relying on new and emerging technologies.

Stefanie Lindstaedt is head of the Institute for Knowledge Management and Knowledge Technologies at Graz University of Technology (Austria) and Scientific Director of the Know-Center in Graz, Austria's competence center for knowledge-based systems. Her research focuses on context-aware knowledge services that combine the power of Web 2.0 approaches with machine learning methods to augment semantic technologies in order to support individual, community, and organisational learning. Before coming to Graz she held various responsible positions such as research project manager at Daimler (Chrysler) Research in Ulm (Germany) and product manager for web-globalisation services at GlobalSight in Boulder (USA).

1 Introduction

Technology Enhanced Learning (TEL) is undergoing a significant shift in paradigm towards more data driven systems that will make educational systems more transparent and predictable. Data science and data-driven tools will change the evaluation of educational practice and didactical interventions for individual learners and educational institutions. We summarise these developments under the keyword *dataTEL* that stands for 'Data-Supported Technology-Enhanced Learning'.

Through the increasing application of learning management systems (LMS), intelligent tutoring systems (ITS), personal learning environments (PLE), Web 2.0 developments, and the EPUB3 standard[1] the collection of large amounts of learning activity data becomes conventional. Although the TEL domain already stores data in their e-learning environments automatically, it typically lacks research on TEL datasets. The unused educational datasets offer an unexploited potential for the evaluation of educational interventions such as new learning settings, learning theories, and learning technologies.

A big advantage of the new dataTEL research is the possibility to collect educational data without any additional efforts like filling a questionnaire or conducting a pre- and post-test assessment; instead the learning activity data can be collected, analysed and reported on demand and therefore offers a new quality for the investigation of the phenomena 'learning'. The collected learning activity data can be presented in real-time and according to the needs of different educational stakeholders like students, teachers, parents, institutions and decision makers. The representation of the learning activity data can support reflection and prediction processes of all stakeholders (students, teachers, and management) and therefore decrease the time that is needed to evaluate the effects of educational contents, scenarios, technology and more general educational interventions.

Furthermore, the increasing research on educational datasets will create more transparent, mutually comparable, trusted and repeatable experiments that lead to evidence-driven knowledge with approved indicators for theories in education and in particularly in TEL. Thus, educational datasets have the potential to facilitate a theory for TEL by providing constant evaluation and experiment settings to analyse learning. It can provide these constant evaluation settings for the whole spectrum of learning from formal to informal learning. Therefore, the educational datasets extend the methodological and empirical approaches to analyse learning and especially technology-enhanced learning

that is still dominated by design-based research approaches, simulations, and field studies. The available educational data introduce an additional research paradigm to the TEL field that has the potential for new insights into learning processes by making so far invisible patterns in the data visible to researchers, educators and the learners.

Based on the outcomes of the *1st dataTEL workshop on Datasets for Technology Enhanced Learning* at the Alpine Rendez-Vous 2011 in France and an open call for submissions, this special issue aims to give an overview of the achievements and developments in this new research field within TEL. This issue clearly demonstrates the new research paradigm by providing insights in the form of extended versions of the workshop papers and related research activities.

Based on the contributions of the participants the workshop organisers identified four most pressing research topics of the workshop and clustered the workshop contributions accordingly. The topics were:

1 Datasets from learning object repositories and web content

2 Evaluation of data-supported tools in TEL

3 Privacy and data protection for educational datasets

4 Data supported learning examples.

During and after the workshop the group further worked on these topics and finally defined four Grand Challenges for dataTEL research for the upcoming future.

In the next section we present the four Grand Challenges that emerged from the research topics. Each of the challenges was structured according to four criteria: (a) problem description, (b) needed actions to address the challenge, (c) feasible timeframe to overcome the challenge, and (d) measurable progress and success indicators. A detailed description of the Grand Challenges, their related research questions and funding opportunities can be found online in the open archive of the STELLAR scientific portal.[2]

2 Grand Challenges for data-supported education

2.1 Grand Challenge 1: a generic framework to share, analyse, and reuse educational datasets

(a) *Grand Challenge description*:

The increased application of LMS, e-portfolios, and PLEs in schools and higher education institutions produces large amounts of educational data. But, although these e-learning environments store educational data automatically, exploitation of this data for new learning services and advanced research on the phenomena ‘learning’ is still very limited. Thus, there is an unused opportunity for the evaluation of learning theories, the development of future learning applications, and the evaluation of didactical concepts and educational interventions. A generic framework to share educational datasets for research purpose is needed to gain new and comparable findings on these datasets.

(b) *Needed actions to overcome the Grand Challenge Problem*:

- Data ownership and access rights are challenging because the LMS and PLE systems are collecting educational data and the current assumption is that this data belongs to them. However, who exactly holds the ownership of the data created by the students and what can be done with it is still unresolved.
- Data policies (licences) that regulate how different users can use, share, and reference certain datasets are very limited available in educational institutes. It needs to be explored how existing licenses like Creative Commons can be applied to educational data as a standard to grant permissions to datasets.
- There is a lack of common dataset formats like suggested from the CEN PT Social Data group[3] and a standardised ways to document such datasets so that others can make proper use of it.
- Standardised methods are needed to anonymise and pre-process educational data according to privacy and legal protection rights.

(c) *Timeframe for the Grand Challenge Problem*:

Five to eight years will be needed to overcome the current situation and achieve more sharable datasets. For learning resource data there are already standards like LOM and Dublin Core. For learning activities and tracking data there is no standard available (apart from very generic formats such as XML – which does not guarantee that data can be reused).

(d) *Measurable progress and success indicators*:

- An increasing amount of publicly available datasets and research articles that are based on shared datasets
- The availability of data or privacy policies at educational providers
- More data-driven tools at educational providers
- A common dataset format.

2.2 Grand Challenge 2: improve course completion and reduce drop-outs through data-driven technologies

(a) *Grand Challenge description*:

A challenging problem for educational institutes is the high drop-out rate especially in online and distance education settings. Isolated study situations at the home place cause a significant amount of students to withdraw from their studies. The research on data supported educational tools can contribute to decrease the drop-out rate by disseminating its research outcomes for the development of new support tools for teachers and students that offer relevant information at the right time. Examples of such tools are drop-out analyser, peer and content recommender systems, and reflection tools. A drop-out analyser can support a tutor of a LMS course by marking students that are likely to drop-out. The tutor can than make an intervention and contact the student personally. Recommender systems can support the students to stimulate them to team-up with related peers or focus on related learning activities.

(b) *Needed actions to overcome the Grand Challenge Problem*:

- Customise existing information retrieval technologies for learning settings
- Employ data supported tools in real-life scenarios and run long time studies
- Develop suitable evaluation criteria for different kind of data supported tools like recommender systems.

(c) *Timeframe for the Grand Challenge Problem*:

There is more than ten years of research on educational data mining and TEL recommender systems. Prototypical solutions can be implemented within a year. More challenging is the evaluation of the tools in long time studies with a runtime of two to three years. For the further development of such systems publicly available educational datasets are needed to evaluate and compare different approaches and to gain a solid body of knowledge on which technique has positive effects on a certain learning situation (five to eight years).

(d) *Measurable progress and success indicators*:

Measurable progress and success indicators are depending on the applied type of data driven tools (peer recommender system, drop-out analyser). For this Grand Challenge a significant decrease of the drop-out rate within an educational institution would be a strong evidence to value the impact of such a system. A challenging issue will be to isolate the effect of decreasing drop-out rates only to a data tool as most educational institutes permanently improve their educational services and various side effects need to taken into account.

2.3 Grand Challenge 3: accurate handling of educational data

(a) *Grand Challenge description*:

Europe's education systems suffer from decreasing amount of teachers and the request to increase the amount of high-educated students in a short time period. As a consequence there is less time available for the individual support of students, thus the teaching quality decreases. On the other hand, the education systems are increasingly based on electronic systems like LMS. With the increase in available educational data, the application of information retrieval technologies becomes valuable to create new services for the educational stakeholders. The combination of educational data and information retrieval techniques also known as Learning Analytics (LA) will become a powerful means in educational practice and student guidance. LA promises the educational field to reduce delivery costs, create more effective learning environments and experiences, accelerate competence development, and increase collaboration between students and teachers. But LA also has barriers and limitations among these are issues of privacy and data protection that need to be addressed by policy guidelines. Additional, challenges arise with respect to data surveillance[4] (social sorting, cumulative disadvantages) and its ethical implications.

(b) *Needed actions to overcome the Grand Challenge Problem*:

- In order to discuss and improve the above-mentioned situation a new vocabulary needs to be accomplished to discuss privacy, data protection and surveillance issues. For instance, what are better terms to express concepts like ownership and access control, when in digital systems replication and distribution is so easy that the concepts have no traction.
- Research is needed on how existing privacy and transparency solutions can be integrated in dataTEL practice. Further, research is desirable on how state-of-the-art security solutions can be used to secure large educational datasets.
- Right treatment of personal data – data awareness in the society. Educational programmes should not be limited to teaching individuals when to reveal or conceal their data, but also to increase their awareness with respect to large datasets, surveillance practices, and related problems.
- More stakeholder studies are necessary to understand the complex requirements with respect to privacy, data protection, and surveillance in dataTEL.
- The issues around privacy, data protection, and surveillance need to be addressed from the beginning of the research and not as an add-on. Methodologies and guidelines that support this vision need to be developed to support privacy and ethical practices.
- There needs to be research on how to bridge and improve communication between dataTEL researchers and ethical boards with respect to advances in technologies and research and the related privacy, data protection, and surveillance concerns that arise with the data.
- Policies have to be defined to avoid unethical data mining research.

(c) *Timeframe for the Grand Challenge Problem*:

The first four activities can be addressed in a time frame of two to three years because they mainly require the application or translation of existing examples or solutions from other domains to the educational field. The activities 5 to 7 will require a longer timeframe (three to five years) as they can only be developed out of the experiences with the activities 1 to 4.

(d) *Measurable progress and success indicators*:

Measureable progress and success indicators are an increasing amount of ethical boards in LA units at educational organisations, an increase of privacy and data protection statements in research projects as well as between educational providers and the students, and an increase of the integration of data and privacy competences in job profiles at educational providers.

2.4 Grand Challenge 4: make data supported information systems an effective tool for educational practice

(a) *Grand Challenge description*:

In order to make data supported information systems an effective tool for educational practice, various limitations and hurdles in technology, privacy and education need to be addressed. It is important to realise that data supported tools work with computational results that are not easy to understand and need to be presented in an easy way (e.g. by visualisations) to address the daily practice of the educational stakeholders. It is crucial to interpret the presented outcomes in a correct manner to take the right follow-up activities that can lead to improved learning. Therefore, the interpretation of educational data and its related tools requires new competences to deal with the outcomes like statistical knowledge, critical thinking, privacy awareness and ethical competences.

(b) *Needed actions to overcome the Grand Challenge Problem*:

- Developing new data driven tools that are easy to understand
- Make new real time data tools available as test applications
- Identify suitable algorithms and map them to certain datasets and learning purposes
- Integrate statistical, critical thinking, privacy awareness and ethical competences into the teacher education programmes.

(c) *Timeframe for the Grand Challenge Problem*:

The development of the data tools has already started, on every scientific event new data driven tools are presented. Systems like Mendeley and open access journals show us already the future of academic work. The training of the new competences for teachers and students in the primary and secondary education level will take five to ten years with having many different levels in the EU partner countries.

(d) *Measurable progress and success indicators*:

- Ethical guidelines how to react on results of data supported learning tools
- Courses at educational providers that train statistics and the interpretation of learning analytic data
- An increasing amount of data tools for different educational stakeholders.

3 Selected papers

In this special issue, several promising research efforts are presented that already start to address the Grand Challenges outlined above.

3.1 Grand Challenge 1

Grand Challenge 1 is mainly addressed by three papers that are presented in the following paragraphs.

Reffay, Betbeder and Chanier describe results of the Mulce project that provides a platform for sharing corpora among researchers in the TEL field. The platform defines a common format for structuring datasets to enable exchange and reuse. In addition, access to datasets that have been captured in real-life settings is provided to researchers in this field.

Jack, Hristakeva, de Zuniga and Granitzer present research on the collection and sharing of a large dataset that was collected by Mendeley – a scientific platform to enable sharing of research papers. The dataset was contributed to the first dataTEL system market that was jointly organised by the 4th ACM Conference on Recommender Systems and the 5th European Conference on Technology Enhanced Learning (EC-TEL 2010) in September 2010. The dataset as well as the Mendeley API to retrieve data on-the-fly are discussed, as well as the potential usefulness of these data for dataTEL research purposes.

Stoitsis, Manouselis and Sanchez-Alonso describe research that has been conducted with data collected from the Organic.Edunet portal, a second dataset that was submitted to the dataTEL market place. The portal provides access to a large collection of learning resources. Data that has been collected from the actual usage with these resources is used by the authors to analyse requirements and guidelines with respect to multilinguality. More specifically, the authors use the dataset to analyse whether the linguistic profile or a portal (i.e. language of interface, of metadata records and of learning resources) may affect the number of the users that is attracted and their search behaviour. In addition, guidelines for information that datasets could store to facilitate multilingual learning analytics are presented.

3.2 Grand Challenge 2

Grand Challenge 2 is addressed by one article in the workshop but there are many articles in the literature that suggest standardisation approaches for the evaluation of different data tools. A good overview can be found in the book *Recommder Systems for Learning*, Springer 2012.[5]

Mulwa, Lawless, O'Keeffe, Sharp and Wade discusses challenges related to the evaluation of the overall performance of personalised learning systems. The paper proposes a novel recommender framework to identify appropriate evaluation methods, metrics and criteria. The system recommends evaluation approaches to software developers and end users of personalised learning systems, based on an extensive analysis of evaluation studies.

3.3 Grand Challenge 3

Grand Challenge 3 is mainly represented by a paper from *Herder and Kawase* who discuss measures and techniques for addressing privacy issues and strategies for motivating people to share their data. More specifically, the authors report on their experiences in building an anonymised dataset of web usage data in the Web History Repository. In addition, guidelines for researchers who intend to recruit participants for creating datasets are discussed.

3.4 Grand Challenge 4

Grand Challenge 4 is addressed by two papers that present two different approaches for dataTEL research. Both approaches share that the research would have been not possible without the datasets and they provide new insights and services for learners to improve their comptences.

Garcia, Pardo, Kloos, Niemann, Scheffel and Wolpers propose the use of visualisation techniques to enable insight into the analytics process. A set of visualisations is used to gain insights into data that is tracked from a virtual machine and that captures the use of different tools and resources in the learning process.

Sie, Drachsler, Bitter-Rijpkema and Sloep present a data supported peer recommender system that suggests future co-authors for scientific article writing based on a Dspace dataset. The recommendation approach is based on network information (betweenness centrality) and author (keyword) similarity. A similar approach could also be applied to written essays in the class setting or to written reports in a company set-up.

4 Final remarks

We believe that the present issue provides several interesting and relevant papers that quite timely describe current developments and achievements. They clearly demonstrate the new data paradigm by providing insights in the form of extended versions of the workshop papers and related research activities.

All the bibliography covered by this special issue is available in an open group created at the Mendeley research platform[6] and will continue to be enriched with additional references. We would like to encourage the reader to sign up for this group and to connect to the community of people working on these topics, gaining access to the collected blibliography but also contributing pointers to new relevant publications within this very fast emerging domain. We hope that you will enjoy reading this special issue.

Notes and References

1 See http://idpf.org/epub/30

2 See http://oa.stellarnet.eu/open-archive/browse?resource=6756_v1

3 See http://sites.google.com/site/camschema/home

4 Data surveillance refers to the process which individualises each member of the population (or a group), and permits the observation and recording of each individual's activities, then collates these individual observations across the population. From these conglomerated observations, statistical norms are produced relating to any of a multitude of characteristics. These norms are then applied back to the subjected individuals, who are categorised and perhaps acted upon, either with gratification or punishment, according to their relation to the produced norm (Phillips, Privacy Policy and PETs, 2004).

5 Manouselis, N., Drachsler, H., Verbert, K. and Duval, E. (2012) *Recommender Systems for Learning*, Springer, Berlin.

6 See http://www.mendeley.com/groups/1919971/celstec-learning-analytics-and-data-supported-education-datatel/