Fostering Analytics on Learning Analytics Research: the LAK Dataset

Davide Taibi
Institute for Educational Technologies
National Research Council of Italy
Via Ugo La Malfa 153, Palermo
Italy
davide.taibi@itd.cnr.it

Stefan Dietze
L3S Research Center
Appelstr. 9a
30167 Hannover
Germany
dietze@l3s.de

ABSTRACT

This paper describes the Learning Analytics and Knowledge (LAK) Dataset, an unprecedented collection of structured data created from a set of key research publications in the emerging field of learning analytics. The unstructured publications have been processed and exposed in a variety of formats, most notably according to Linked Data principles, in order to provide simplified access for researchers and practitioners. The aim of this dataset is to provide the opportunity to conduct investigations, for instance, about the evolution of the research field over time, correlations with other disciplines or to provide compelling applications which take advantage of the dataset in an innovative manner. In this paper, we describe the dataset, the design choices and rationale and provide an outlook on future investigations.

Categories and Subject Descriptors

H.3.4 [**Semantic Web**], I.2.4 [**Ontologies**]

General Terms

Algorithms, Documentation, Design, Experimentation, Standardization.

Keywords

Learning Analytics, Data, Linked Data, Semantic Web, Educational Data Mining

1. INTRODUCTION

As part of an international team of research practitioners consisting of the Society for Learning Analytics Research (SoLAR)[1], ACM[2], the LinkedUp project[3], the Educational Technology Institute of the National Research Council of Italy (CNR-ITD), we have released an unprecedented resource for the Learning Analytics and Educational Data Mining. In order to allow the computational analyses of research publications in the emerging Learning Analytics field, we have published in a machine-readable format a comprehensive set of scientific papers in the field of learning analytics and educational data mining. This can serve as input for studies into scientometrics, investigations into the evolution of the overall discipline or correlations with other fields.

The LAK dataset provides access to documents already available online in unstructured form but also to research works not publicly accessible before at all. The collection includes the proceedings of the International Educational Data Mining Society[4], as well as the Journal of Educational Technology and Society a special issue on Learning Analytics[5]. In both cases full text has been freely available already but in unstructured format. Content which has previously been accessible to subscribers only includes the "Proceedings of the ACM International Conference on Learning Analytics and Knowledge" edited by ACM. In the framework of our initiative, ACM is providing freely its ACM Digital Library in the Learning Analytics field solely for research purposes.

The following table describes in details the set of papers that have been collected, processed and exposed in a structured form in the Learning Analytics and Knowledge (LAK) dataset:

Table 1 : Papers included in the LAK dataset

Publication	**# of papers**
Proceedings of the ACM International Conference on Learning Analytics and Knowledge (LAK) (2011-12)	66
The open access journal Educational Technology & Society special issue on "Learning and Knowledge Analytics": Educational Technology & Society (Special Issue on Learning & Knowledge Analytics, edited by George Siemens & Dragan Gašević), 2012, 15, (3), pp. 1-163.	10
Proceedings of the International Conference on Educational Data Mining (2008-12)	239
Journal of Educational Data Mining (2008-12)	16

[1] www.solaresearch.org/

[2] http://acm.org/

[3] http://linkedup-project.eu/

[4] http://www.educationaldatamining.org/proceedings

[5] http://www.ifets.info/issues.php?id=56

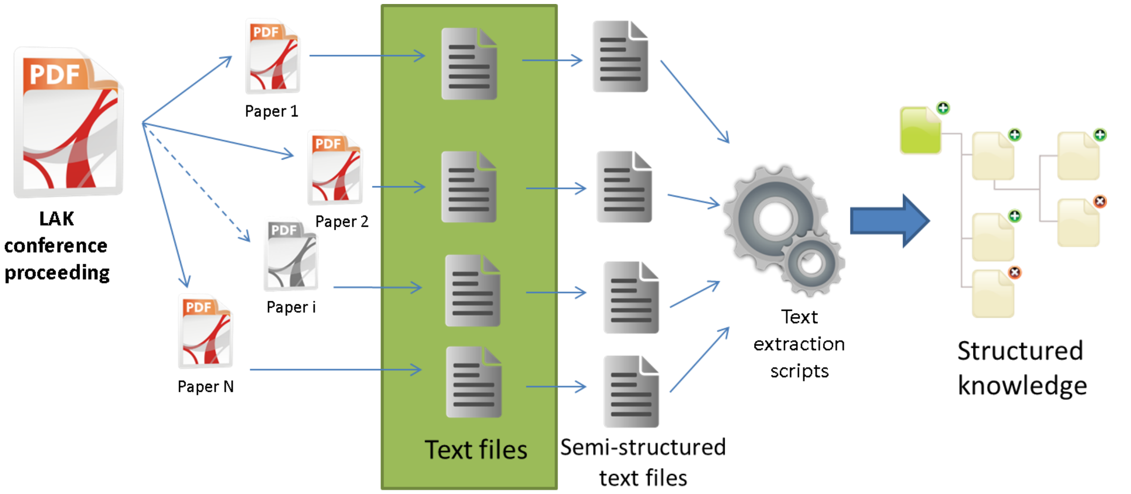

Figure 1: Knowledge extraction process

2. FROM SCHOLARLY PAPER TO STRUCTURED DATA

2.1 The extraction process

In order to process and analyze the set of unstructured journals and conferences papers data was transformed into structured data. While each conference proceeding is available on the Web in PDF format, but each collection has its own structure. Even if in some cases the most used format is the ACM template[6], papers not always comply with it entirely, calling for some specifically adapted extraction mechanisms. The overall knowledge extraction process is composed of three main steps:

1. Transforming PDF documents to plain textual representation.

2. Cleaning up and consolidation of the textual information.

3. Extracting structured data from text.

In the first step the PDF file containing the proceeding of a conferences, or the papers of a journal is split up in order to have one document for each paper. Then each PDF file has been elaborated with pdf2text tool in order to have a textual representation for each paper.

In the second step the text files are elaborated in order to transform them in a partially structured format that can be elaborated automatically. In particular at this step tables and figures are removed from the paper, maintaining their captions, that can be useful for text mining processing, footnotes have been also removed from the text, while bulleted or numbered list have been organized using an homogeneous format.

As part of the third step, text files are being processed in order to extract from them the most important sections of the document. Regarding the authors, their *name, affiliation, country* are represented using the FOAF ontology. For each paper the following information are collected: *title, authors, keywords, abstract, full-text* and its relationship with the type of publication or event, (journal or conference proceedings).

It is important to note that beside the common metadata for the learning analytics papers such as: title, abstract, authors and affiliation also the full text of the papers is stored in the dataset. At this stage the full text has been stored without considering its separation in paragraphs and sections, however the elaboration performed at step number 2 has also identified the titles and paragraphs of sections and subsections, thus providing the basis for analyzing full text with further granularity in next versions of the LAK dataset. The referenced papers are also extracted but are not made available in the LAK dataset in this version of the dataset .

2.2 The schema

The schema used to describe the papers in the dataset is based on two established schemas: the Semantic Web Conference (SWC) ontology[7] (already used to describe metadata about publications from the Semantic Web conferences and related events[8]) and the Linked Education schema[9]. The Linked Education schema has been developed to represent and catalog both educational and educational related datasets, which are datasets not specifically created for education but that can be used in an educational context. The schema has been used to annotate datasets and resources as part of an integrated dataset[10] which contains educationally relevant resources such as : *LinkedUniversities[1]* and the *mEducator Educational Resources* [4] with their Open Educational Resources and materials explicitly related to education, as well as implicitly educationally relevant datasets such as *BBC Programmes*[3], *ACM Library Metadata[11]* and *Europeana [2]* datasets. The main entities collected in the LAK dataset are paper authors, institutions and papers, related to the

[6] http://www.acm.org/sigs/publications/proceedings-templates

[7] http://data.semanticweb.org/ns/swc/ontology

[8] http://data.semanticweb.org/

[9] http://data.linkededucation.org/ns/linked-education.rdf

[10] http://linkedup.l3s.uni-hannover.de:8880/openrdf-sesame/repositories/linked-learning-selection?query

[11] http://acm.rkbexplorer.com/

learning analytics area. Authors and institutions have been represented using respectively the classes Person and Organization of the FOAF ontology, while to represent papers, the class *InProceedings* of the SWRC ontology has been used. The LAK Dataset, at the time of writing, includes 779 authors, connected to 295 institution, and 315 posters, abstract, short and full papers.

2.3 Access Methods

In order to support different access method for the data, the resources of the LAK datasets have been published in different formats:

- A dump file in zipped RDF/XML file format can be directly downloaded from the SoLAR research web page.
- A version of the dataset in a format that can be elaborated through the R statistic software have been provided[12]
- A Linked Data endpoint with a public SPARQL endpoint has been developed in order to provide access to structured RDF metadata according to LOD principles.[13]

The following SPARQL query[14] on the LAK dataset can be used to extract the full text of all 2011 papers (LAK 2011, and EDM 2011 conferences) in .srx format (XML file which can be opened in any text editor):

```
PREFIX led:<http://data.linkededucation.org/ns/linked-education.rdf#>
PREFIX swrc:<http://swrc.ontoware.org/ontology#>
SELECT ?paper ?fulltext WHERE { ?paper led:body ?fulltext . ?paper
swrc:year ?year . FILTER (?year = “2011”) }
```

On the SoLAR Website some useful examples for querying the SPARQL endpoint[15] of the LAK dataset have been reported[16]. The example queries allow users, for instance, to retrieve: the papers co-authored by two selected authors; all papers published in both EDM and LAK conferences by the authors affiliated to an institution.

3. LAK DATA CHALLENGE

Beyond merely publishing the data, we are actively encouraging its innovative use and exploitation as part of a public *LAK Data Challenge*[17] sponsored by the European Project LinkedUp. An initial competition is co-located with the ACM LAK13 Conference, Leuven, Belgium (April 2013)[18]. The challenge is revolving around the overall question on what insights can be gained from analytics on the LAK corpus about the general discipline of Learning Analytics and its connection to other fields. How can we make sense of this emerging field’s historical roots, current state, and future trends, based on how its members report and debate their research? Challenge submissions should exploit the LAK Dataset by covering one or more of the following, non-exclusive list of topics:

- Analysis & assessment of the emerging LAK community in terms of topics, people, citations or connections with other fields
- Innovative applications to explore, navigate and visualise the dataset (and/or its correlation with other datasets)
- Usage of the dataset as part of recommender systems

4. CONCLUSIONS

The LAK Dataset has been a first starting point to allow the analysis of leaning analytics as an emerging research discipline and its definition and evolution. Along with the growth of the research works in learning analytics and related fields, we intend to expand the dataset by adding new research publications. In addition, while the dataset currently contains plain metadata and full text of the research publications, it is envisaged to extract and add additional data about contained entities and topics, to provide simple means for assessing, exploring and navigating the data.

5. ACKNOWLEDGMENTS

This work is partly funded by the European Union under FP7 Grant Agreement No 317620 (LinkedUp).

6. REFERENCES

[1] Fernandez, M., d'Aquin, M., and Motta, E. 2011. Linking Data Across Universities: An Integrated Video Lectures Dataset. *In Proceeding of the 10th International Semantic Web Conference (ISWC 2011)*, 23 - 27 Oct 2011, Bonn, Germany.

[2] Haslhofer, B., Isaac. A. 2011. data.europeana.eu - The Europeana Linked Open Data Pilot. *In Proceeding of the International Conference on Dublin Core and Metadata Applications* (DC 2011).

[3] Kobilarov, G., Scott T., Raimond Y., Oliver S., Sizemore C., Smethurst M., Bizer C., Lee R. 2009. Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Conections. *In Proceedings of the 6th European Semantic Web Conference* (ESWC2009).

[4] Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P., Bratsas, C. and Woodham, L. 2011. Connecting Medical Educational Resources to the Linked Data Cloud: the mEducator RDF Schema, Store and API, *in Linked Learning 2011, Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age*, CEUR-WS, Vol. 717, 2011

[12] http://www.r-project.org/

[13] http://data.linkededucation.org/openrdf-sesame/repositories/lak-conference?query=[your sparql query]

[14] http://data.linkededucation.org/openrdf-sesame/repositories/lak-conference?queryLn=SPARQL&query=PREFIX%20led%3A%3Chttp%3A%2F%2Fdata.linkededucation.org%2Fns%2Flinked-education.rdf%23%3E%0APREFIX%20swrc%3A%3Chttp%3A%2F%2Fswrc.ontoware.org%2Fontology%23%3E%0A%0Aselect%20%3Fpaper%20%3Ffulltext%20where%20%7B%3Fpaper%20led%3Abody%20%3Ffulltext%20.%20%3Fpaper%20swrc%3Ayear%20%3Fyear%20.%20FILTER%20%28%3Fyear%20%3D%20%222011%22%29%20%7D&infer=true

[15] http://data.linkededucation.org/openrdf-sesame/repositories/lak-conference

[16] http://www.solaresearch.org/resources/lak-dataset/sparql-queries/

[17] http://www.solaresearch.org/events/lak/lak-data-challenge/

[18] http://lakconference2013.wordpress.com/