Facilitating Scientometrics in Learning Analytics and Educational Data Mining - the LAK Dataset

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Davide Taibi[a,*], Stefan Dietze[b] , Mathieu d'Aquin[b]
[a]*National Research Council of Italy, Institute for Educational Technologies, via Ugo La Malfa 153, 90146 Palermo, Italy*
[b]*L3S Research Center, Appelstrasse 9a, 30167 Hannover, Germany*
[c]*Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, UK, MK7 6AA*

Abstract. The Learning Analytics and Knowledge (LAK) Dataset represents an unprecedented corpus which exposes a near complete collection of bibliographic resources for a specific research discipline, namely the connected areas of Learning Analytics and Educational Data Mining. Covering over five years of scientific literature from the most relevant conferences and journals, the dataset provides Linked Data about bibliographic metadata as well as full text of the paper body. The latter was enabled through special licensing agreements with ACM for publications not already available through open access. The dataset has been designed following established Linked Data pattern, reusing established vocabularies such as the SWRC ontology or BIBO and providing links to established schemas and entity coreferences in related datasets. Given the temporal and topic coverage of the dataset, it facilitates scientometric investigations, for instance, about the evolution of the field over time, or correlations with other disciplines, what is documented through its usage in a wide range of scientific studies and applications.

Keywords: Learning Analytics, Educational Data Mining, Linked Data

1. Introduction

While there exist a wealth of datasets containing bibliographic metadata, such as ACM[1] or DBLP[2], such corpora provide RDF data covering bibliographic metadata such as authors, affiliations and publication metadata, but - with positive exceptions such as the Semantic Web Journal - usually lack direct access to the content of the publication. This is partly due to licensing constraints as well as lack of sufficient open access publications. Essentially, it hinders researchers from carrying out scientometric investigations or to deeply investigate the evolution of scientific disciplines, topics or researchers over time.

The Learning Analytics and Knowledge (LAK) Dataset[3] represents an unprecedented corpus which exposes a near complete collection of bibliographic resources for a specific research discipline, namely the connected areas of Learning Analytics (LA) and Educational Data Mining (EDM). Covering over five years of scientific literature from the most relevant conferences and journals in these disciplines, the

[1] http://datahub.io/dataset/rkb-explorer-acm
[2] http://datahub.io/dataset/l3s-dblp
[3] http://lak.linkededucation.org

[*]Corresponding author. E-mail: davide.taibi@itd.cnr.it

dataset provides Linked Data about bibliographic metadata as well as full text for all publications. The latter was enabled through special licensing agreements with ACM for publications not already available as open access.

The dataset is published and maintained with support of the LinkedUp project[4], the Society for Learning Analytics Research (SoLAR)[5], ACM[6], the L3S Research Center [7] and the Institute for Educational Technology of the National Research Council of Italy (CNR-ITD)[8], with the main goals being (i) facilitating scientific analysis of the LA/EDM communities and (ii) improving access to scientific literature in the field of LA/EDM. Beyond merely publishing the data, the use and exploitation of the dataset is actively encouraged by means of the annual LAK Data Challenge, which has led to the emergence of an increasing number of applications and studies.

2. The LAK Dataset

The LAK Dataset provides scientific publications from major conferences in the the LA and EDM field in a machine-processable format (RDF). While we also offer regularly updated dumps (RDF/XML, N-Triples and R), here we specifically discuss the RDF dataset and SPARQL endpoint, accessible as described in Table 1.

Table 1 LAK Dataset facts

Name	LAK Dataset
Dataset Home	http://lak.linkededucation.org
Example resource	http://data.linkededucation.org/resource/lak/conference/lak2013/paper/93
SPARQL endpoint	http://data.linkededucation.org/request/lak-conference/sparql
Dump (RDF/XML)	http://lak.linkededucation.org/lak/LAK-DATASET-DUMP.rdf.zip
Dump (R)	http://crunch.kmi.open.ac.uk/people/~acooper/data/LAK-Dataset.RData
Dump (N-Triples)	http://lak.linkededucation.org/lak/LAK-DATASET-DUMP.nt.zip
Publication date	12/12/2012
Last update	30/09/2014
License	Creative Commons Attribution Share-Alike (cc-by-sa)
Datahub.io	http://datahub.io/dataset/lak-dataset

[4] http://linkedup-project.eu
[5] http://www.solaresearch.org/
[6] http://acm.org/
[7] http://www.l3s.de/
[8] http://www.itd.cnr.it/

2.1. Coverage and data sources

The LAK dataset exposes metadata and full text of key academic publications in LA and EDM. Papers are sourced from the two major conferences in each field (ACM Learning Analytics and Knowledge, International Conference on Educational Mining), the two main journals, namely the recently founded Journal of Learning Analytics and the Journal of Educational Data Mining, and the proceedings of the two editions of the LAK Data Challenge held in conjunction with the LAK conferences. Table 2, shows the number of papers included from each source. This collection constitutes a near complete corpus of research works in the areas Learning Analytics and Educational Data Mining.

2.2. Schema/ontology

To ensure wide interoperability of the data, the schema is based in particular on the Semantic Web Conference Ontology[9], which is a well-established schema already used to represent bibliographic metadata, for instance, as part of the Semantic Web Dogfood dataset[10]. Where more expressivity was required, additional elements from related vocabularies are used, including the Bibliographic Ontology BIBO [11], Schema.org, and Nature Publishing Group (NPG) Ontology Terms [12], including the corresponding type alignments.

For each publication the following features are extracted: title, authors, keywords, abstract, text body, references, publication venue (journal/conference proceedings).

Authors and institutions have been represented using respectively the types *Person* and *Organization* of the FOAF ontology, whereas the type *InProceedings* of the SWRC ontology has been used to represent conference papers and the type *Article* of the the BIBO ontology for Journal publications. The property *bibo:content* of BIBO (aligned with *schema:articleBody* as *owl:equivalentProperty)* represents the full text of the papers. The NPG ontology is employed to represent references by means of the class *npg:Citation*, referred to via the *npg:hasCitation* property.

[9] http://ontoware.org/swrc/
[10] http://data.semanticweb.org/
[11] http://bibliontology.com/
[12] http://ns.nature.com/terms/

Table 2 Academic publications presented in the dataset

Publication	# of papers
Proceedings of the ACM International Conference on Learning Analytics and Knowledge (LAK) (2011- 2014)	166
Proceedings of the International Conference on Educational Data Mining (2008-2014)	463
Special issue on "Learning and Knowledge Analytics": Educational Technology & Society, edited by George Siemens & Dragan Gašević), 2012, 15, (3), 1-163.	10
Journal of Educational Data Mining (2009-2014)	29
Journal of Learning Analytics (2014)	16
Proceedings of the LAK data Challenge (2013-2014)	13

By relying entirely on established and frequently used types and properties, we aim for a high reusability of the data. While SWRC had shown a high overlap with the conceptual model of our dataset, it was used as starting point and gradually expanded with additional elements to fully represent the data model of the LAK dataset. Property type associations in our dataset have led to a number of implicit mappings between types, such as *swrc:InProceedings*, *schema:Article* and *bibo:Document* which are further supported and complemented through entity links described in the following section.

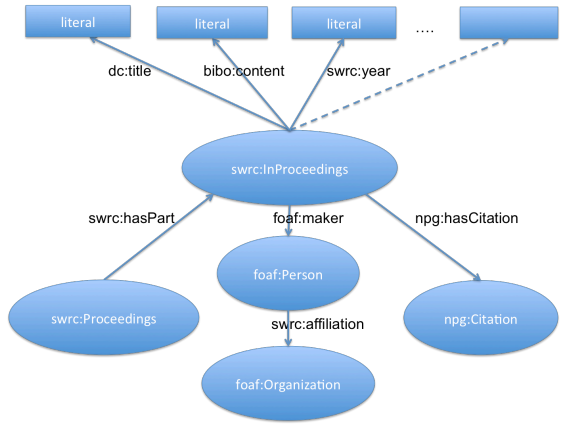

Fig. 1: Key classes and properties used in the LAK dataset (conference proceedings only)

The following table provides a general overview of the number of represented entities per type in the LAK dataset.

Table 3 Entity population in LAK Dataset

Concept	ofType	#
Reference	npg:Citation	7885
Author	foaf:Person	1214
Conference	swrc:InProceedings	652
Paper		
Organization	foaf:Organization	365
Journal Paper	bibo:Article	45
Proceedings Volume	swrc:Proceedings	15
Journal Volume	bibo:Journal	9

Table 4 summarizes the most frequently populated properties.

Table 4 Most frequently populated properties in the LAK Dataset

Domain	Property	Range	#
npg:Citation	rdfs:label	literal	10828
swrc:InProceedings	dc:subject	literal	3392
swrc:InProceedings	foaf:maker	foaf:Person	2199
foaf:Person	foaf:made	swrc:InProceedings	2199
swrc:InProceedings	dc:creator	foaf:Person	2199
foaf:Person	rdfs:label	literal	1583
foaf:Person	foaf:name	literal	1583
foaf:Person	foaf:sha1sum	literal	1341
foaf:Person	swrc:affiliation	foaf:Organization	1293
swrc:InProceedings	bibo:content	literal	698

2.3. Inter-Dataset Links

While bibliographic metadata, particularly about, is widespread in the LOD graph, our interlinking efforts have particularly focused on co-reference resolution across entities such as authors, publications, and organisations. Given the Computer Science-scope of LAK, we have particularly considered the datasets DBLP[13] & Semantic Web Dog Food. While DBLP allows us to link authors to their corresponding representation in a more exhaustive Computer Science bibliographic knowledge base, the Semantic Web Dogfood has been particularly useful to relate equivalent organisations, given its strong overlap with the LAK Dataset with respect to authors' affiliations. All

[13] http://datahub.io/dataset/l3s-dblp

considered datasets complement each other with respect to the schema, i.e. the expressed properties and conceptual model, as well as its population, i.e. the amount of distinct entities actually represented within each dataset. While the LAK Dataset has a high depth with respect to the represented properties and features, even including references and textual body of publications in contrast to most bibliographic databases, it has a fairly narrow scope by focusing entirely on specific CS subjects (Learning Analytics and Educational Data Mining). Coreference resolution of entities, for instance authors, in other more complete Computer Science-related bibliographic knowledge bases provides a more complete view on the work of individual authors or organisations and the CS community as a whole. Similarly, the LAK Dataset complements existing corpora by (a) enriching the limited metadata with additional properties and (b) containing additional publications not reflected in DBLP or the Semantic Web Dogfood, creating a more comprehensive knowledge graph of CS literature as a whole. For instance, in DBLP and Semantic Web Dogfood, LAK publications are not exhaustively represented, references and full text is missing in both cases and, in the case of DBLP, affiliations are not reflected as explicit entities.

While overlap among authors in LAK and Semantic Web Dogfood has been less prominent, the majority of authors could be resolved using DBLP, enabling a broader understanding of the general scientific output of LAK researchers. For establishing coreferences, literals (*foaf:name, dc:title*) of entities in all three datasets have been matched. To improve recall and cater for different representations, some preprocessing was applied to address issues with character codes and distinct naming conventions.

Additional outlinks were created to DBpedia as reference vocabulary. To allow a more structured retrieval and clustering of publications according to their topic-wise similarity we have linked *keywords*, manually provided by paper authors, to their corresponding entities in DBpedia, thereby using DBpedia as reference vocabulary for paper topic annotations. Keywords, i.e. terms, were disambiguated through state of the art NER methods (DBpedia Spotlight), allowing to link for instance keywords such as “Game-based learning” or "educational gaming" to corresponding DBpedia entities such as *http://dbpedia.org/resource/Educational_game*, as in the LAK paper entities *http://data.linkededucation.org/resource/lak/conference/edm2014/paper/580*.

The following figure depicts the total amount of links of resolved respectively enriched LAK entities.

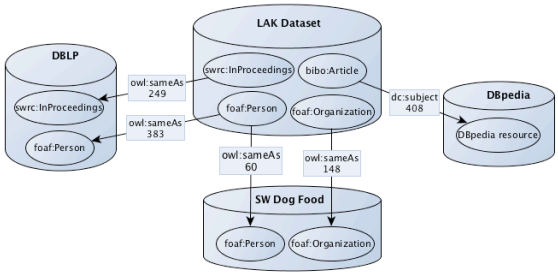

Fig. 2 : Links in the LAK Dataset

With respect to inlinks, the dataset is referenced by the LinkedUp catalog [14] and the majority of its resources are referenced by the Linked Dataset Profiles[15] dataset, further described in [7]. Additional inlinks might have been generated by the works described in [8][8].

3. Query and exploration

The purpose of the dataset is two-fold, (i) facilitating access to scientific literature in LA & EDM and (ii) enabling scientific analysis of the research community, in particular its evolution, interaction pattern and correlation with other communities. Some example queries[16] are reported below.

The interlinks of the LAK dataset with external datasets support federated queries, combining data about the same entity spread across different sources, for instance, papers, authors and properties in LAK, SW Dogfood and DBLP for one specific academic institution. At the same time, term-disambiguation with DBpedia facilitates more precise, entity-based queries, for instance, by using disambiguated DBpedia entities when querying for specific topics (Listing 1).

```
1  PREFIX dc:<http://purl.org/dc/elements/1.1/>
2  PREFIX dbpedia:<http://dbpedia.org/resource/>
3
4  select distinct ?papers ?subject where {
5   ?paper dc:subject dbpedia:Educational_game .
6   ?paper dc:subject ?subject .
```

[14] http://data.linkededucation.org/linkedup/catalog

[15] http://data.l3s.de/dataset/linked-dataset-profiles

[16] Additional queries available: http://solaresearch.org/initiatives/dataset/sparql-queries/]

```
}
```

Listing 1: Retrieving papers covering related topics (sharing same DBpedia entities)

The following example shows a federated query executed across the LAK dataset and the DBLP dataset. In this query, the information about a specific paper of the LAK dataset has been completed with additional data (DOI, reference to bibsonomy) included in DBLP.

```
PREFIX owl:<http://www.w3.org/2002/07/owl#>

select ?dblp ?p ?o where {

<http://data.linkededucation.org/resource/lak/
conference/lak2012/paper/14> owl:sameAs ?dblp

 SERVICE <http://dblp.l3s.de/d2r/sparql>
 {
    ?dblp ?p ?o
 }
}
```

Listing 2: Federated query retrieving bibliographic data related to one paper from DBLP and LAK-Dataset

Listing 3a shows a query to extract statistics on the number of papers per authors over years is reported:

```
PREFIX foaf:<http://xmlns.com/foaf/1.0/>
PREFIX swrc:<http://swrc.ontoware.org/ontology#>

select distinct ?author ?years COUNT(?years)
where {
 ?paper a swrc:InProceedings .
 ?paper swrc:year ?years .
 ?author a foaf:Person .
 ?paper foaf:maker ?author .
} ORDER BY ?author ?years
```

Listing 3a: Total amount of papers per authors/years.

And the following one reports statistics on the number of papers per organization over years:

```
PREFIX foaf:<http://xmlns.com/foaf/1.0/>
PREFIX swrc:<http://swrc.ontoware.org/ontology#>

select distinct ?organization ?years
COUNT(?years) where {
 ?paper a swrc:InProceedings .
 ?paper swrc:year ?years .
 ?paper foaf:maker ?author .
 ?author swrc:affiliation ?organization .
} ORDER BY ?organization ?years
```

Listing 3b: Total papers per organization & year.

Listing 4 shows a query to retrieve influential publications in the LA field by selecting the most cited papers.

```
PREFIX npg:<http://ns.nature.com/terms/>
PREFIX swrc:<http://swrc.ontoware.org/ontology#>

select distinct ?reference
COUNT(*) AS ?count
where {
 ?paper a swrc:InProceedings .
 ?paper npg:hasCitation ?reference
} ORDER BY DESC(?count) LIMIT 10
```

Listing 4: Retrieving influential publications by means of the most cited papers.

4. Creation, maintenance & sustainability

The knowledge extraction process implemented to transform unstructured publications into structured data is composed of three main steps:

1. Transforming PDF to plain textual representation.
2. Pre-processing, clean-up and consolidation of the textual information.
3. Lifting data into RDF schema.

In order to process and analyze the set of unstructured publications (PDF) was transformed with the pdf2text tool into plain text. Given the inherent differences of the structure of papers across the different venues, the extraction had to be tailored to each publication origin. Additional issues arose from papers not complying entirely with the suggested layout, requiring several improvement iterations.

In the second step, extracted text is consolidated and partially structured for further processing. In particular, tables and figures are removed while maintaining their captions for text mining, while bulleted or numbered lists have been organized using a homogeneous format. Finally, data from the semi-structured text documents is lifted into our RDF schema, populating both the metadata as well as the text body related properties. At this stage the full text has been extracted without considering its separation in paragraphs and sections, however the elaboration performed at step 2 has also identified titles and paragraphs of sections and subsections, thus providing the basis for analyzing full text with additional granularity in next versions of the LAK dataset. Literature references are also extracted and are made available in the LAK dataset in order to support scientiometrics based on co-citation networks.

Given the nature of the dataset, we continuously add new publications as these become available, i.e. whenever new proceedings or journal issues of the

reflected series become available. As this process cannot be triggered automatically, we manually observe the community and are receiving notifications from the respective conference and journal chairs as well as SoLAR, being the most central organisation overlooking LA research. The actual insertion process for new publications is straight-forward, given that the processing pipeline is developed and has been refined throughout the last years.

Maintenance is not only carried out at the data or instance level, but also with respect to the actual ontology and its alignment with other vocabularies, e.g. by frequently adding new alignments with emerging vocabularies.

5. Applications &usage

The LAK Dataset has received considerable attention and support from organisations such as the Society of Learning Analytics Research (SoLAR), which also advertises the dataset for its own purposes[17]. Throughout the last years, the dataset has emerged into a central resource for researchers in the LA and EDM field, documented by a variety of research publications which make use of the data. Including the proceedings [8][8], the authors already are aware of 16 scientific publications[18] which make use of the LAK dataset. Events related to the LAK Dataset have been directly supported by the steering board of the LAK conference, most notably the LAK Data Challenge. The LAK Dataset is at the basis of the LAK Data Challenge organised by a team of researchers affiliated with SoLAR, LinkedUp[19] and associated organisations is an annual competition (now in its third year), and previously has been co-located with the ACM LAK conferences in Leuven[20] and Indianapolis[21]. While earlier editions of the challenge were held as workshops or tutorials at ACM LAK, the 2015 edition will be embedded into the main conference tracks. Below, we specifically summarise applications and explorations of the dataset developed by third parties as part of the LAK Data Challenge.

[17] http://solaresearch.org/initiatives/dataset/

[18] Known publications listed at http://lak.linkededucation.org/?page_id=7

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

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

[21] http://lak14indy.wordpress.com/

The challenge is revolving around the overall question on what insights can be gained from analytics on the LAK dataset about the evolution Learning Analytics as a whole or individual topics, researchers or organisation as well as their correlation with other fields.

Given the narrow scope of the data, the variety of the short-listed submissions (so far 13 in total) has been very wide, where Figure 3 gives an overview of the involved author origins.

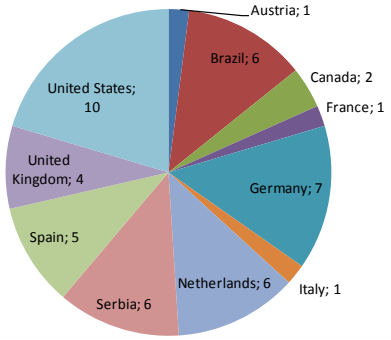

Fig. 3. LAK Data Challenge submissions - authors per country

Applications are further described in the proceedings of the 2013 and 2014 LAK Data Challenge editions [8][8] and are available online[22,23]. Challenge submissions have exploited the LAK Dataset by covering one or more of the following, non-exclusive list of topics:

- *Analysis & assessment* in terms of topics, people, citations or connections with other fields
- *Applications* to explore, navigate and visualize the dataset (and/or its correlation with other datasets)
- Usage of the dataset as part of *recommender systems*

While all submissions are notable and in many cases, combine features from several categories, we would like to emphasize particularly works which have received recognition beyond the challenge, such as "Cite4Me" [4], the analysis of Balachef et al in [5] or near complete scientometric environments such as DEKDIV [6] (depicted in Figure 4), which combine a range of features, such as trending topic analysis, co-citation and collaboration analysis with recommendation approaches, for instance to suggest adequate reviewers and experts.

[22] http://ceur-ws.org/Vol-974/

[23] http://ceur-ws.org/Vol-1137/

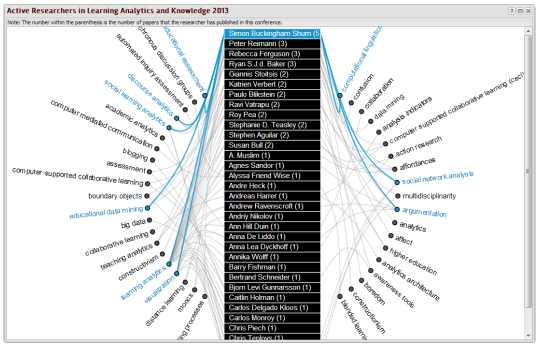

Fig. 4. DEKDIV active LA researchers exploration

6. Discussion & Future Work

According to the 5-star classification[24] of LOD and Vocabulary use (see also [3]) the LAK Dataset qualifies as a 5-star dataset. However, there are known short-comings which the authors are addressing as part of ongoing and future work. While the extraction process is not entirely flawless and, depending on the quality of the source PDFs, had in some cases required manual adjustment, also the automated co-reference resolution had to consider particular draw-backs. For this, we specifically preferred a high precision in favor of recall, to ensure a knowledge graph which is as correct, rather than as complete as possible. We are currently looking into more sophisticated entity interlinking methods which employ similarity measures to match entity features, in order to further increase the linking to related entities in other datasets. In addition, the extraction of references is so far in a preliminary stage, providing the reference in an entirely unstructured manner. Here, as part of upcoming releases, references will be extracted in a more structured format, where features are directly lifted into bibliographic metadata properties.

While the LAK Dataset has a fairly well-defined and somewhat narrow scope, covering only literature in a very specific subdiscipline - i.e. LA and EDM - analysis and correlation with bibliographic information in other sources already now enables interesting investigations and applications [8][8]. However, while the actual text body of publications contains substantial information for a more thorough analysis but is yet still missing from the majority bibliographic LD, we would like to encourage work on similar efforts, i.e. the creation of bibliographic datasets containing both metadata and the actual content. This would allow a more direct processing and analysis of scientific works across disciplines. Furthermore, applying our approach to a wider area could contribute to resolving the gap between unstructured and hard-to process publication formats such as traditional PDFs and structured Linked Data, a topic widely discussed in the Semantic Web community and beyond,

[24] http://www.w3.org/DesignIssues/LinkedData.html

References

[1] H. Drachsler, S. Dietze, E. Herder, M. d'Aquin, D. Taibi: The learning analytics & knowledge (LAK) data challenge 2014. LAK 2014: 289-290

[2] D. Taibi and S. Dietze, (2013), Fostering Analytics on Learning Analytics Research: the LAK Dataset. In: CEUR WS Proceedings Vol. 974, Proceedings of the LAK Data Challenge, held at LAK2013 – 3rd International Conference on Learning Analytics and Knowledge (Leuven, BE, April 2013)

[3] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, Charles Vardeman II, Five Stars of Linked Data Vocabulary Use, Semantic Web, 5 (3), 2014, 173-176.

[4] Nunes, B. P., Fetahu, B., Dietze, S., Casanova, M. A., Cite4Me: A Semantic Search and Retrieval Web Application for Scientific Publications, 12th International Semantic Web Conference (ISWC2013), Sydney, Australia, (2013).

[5] Balacheff, N., Lund, K., Multidisciplinarity vs. Multivocality, the case of "Learning Analytics", LAK '13 Proceedings of the Third International Conference on Learning Analytics and Knowledge

[6] Hu, Y., McKenzie, G., Yang, J.A., Gao, S., Abdalla, A., Janowicz, K., A Linked-Data-Driven Web Portal for Learning Analytics: Data Enrichment, Interactive Visualization, and Knowledge Discovery, In: CEUR WS Proceedings Vol. 1137, Proceedings of the Workshops at the LAK 2014 Conference, co-located with 4th International Conference on Learning Analytics and Knowledge (LAK 2014).

[7] Fetahu, B., Dietze, S., Nunes, B. P., Casanova, M. A., Nejdl, W., A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles, 11th Extended Semantic Web Conference (ESWC2014), Heraklion, Crete, Greece, (2014).

[8] Drachsler, H., Dietze, S., d'Aquin, M., Herder, E., Taibi, D., Proceedings of the LAK Data Challenge 2014, held at LAK 2014, the 4th Conference on Learning Analytics and Knowledge (LAK2014), CEUR Workshop Proceedings, Vol. 1137, 2014.

[9] d'Aquin, M., Dietze, S., Drachsler, H., Herder, E., Taibi, D., Proceedings of the LAK Data Challenge, held at LAK 2013, the Third Conference on Learning Analytics and Knowledge (LAK2013), CEUR Workshop Proceedings, Vol. 974, 2013.