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

Editor(s): Claudia d'Amato, Università degli Studi di Bari, Italy
Solicited review(s): Agnieszka Ławrynowicz, Poznan University of Technology, Poland; Maria Keet, University of Cape Town, South Africa; Vojtěch Svátek, University of Economics, Prague, Czech Republic

Stefan Dietze[a,*], Davide Taibi[b], Mathieu d'Aquin[c]
[a]*L3S Research Center, Appelstrasse 9a, 30167 Hannover, Germany*
[b]*National Research Council of Italy, Institute for Educational Technologies, via Ugo La Malfa 153, 90146 Palermo, Italy*
[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 yet available through open access. The dataset has been designed following established Linked Data pattern, reusing established vocabularies and providing links to established schemas and entity coreferences in related datasets. Given the temporal and topic coverage of the dataset, being a near-complete corpus of research publications of a particular discipline, it facilitates scientometric investigations, for instance, about the evolution of a scientific 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], these usually 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 despite wider calls, for instance at the European level[3], to publish data and scientific output in machine-readable and open formats to facilitate reuse and interoperability.

Such a lack of access to openly licensed and structured research information hinders researchers from

[1] http://datahub.io/dataset/rkb-explorer-acm

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

[3] See Official Journal of the European Union, 2014/C 240/01, 57, (2014), http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:C:2014:240:FULL&from=EN & European Union: Directive 2013/37/EU in Official Journal of the European Union, 56, 2013/L 175/1 (2013), http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L:2013:175:FULL&from=EN

*Corresponding author. E-mail: dietze@l3s.de

carrying out scientometric investigations or to deeply investigate the evolution of scientific disciplines, topics or researchers over time. In particular, for the investigation of the inherent dynamics and the evolution of an entire discipline over time, no dedicated corpus exists which (a) provides bibliographic metadata and full text in a structured and machine-processable format such as Linked Data and (b) covers the near-complete output of a particular research community over its entire existence.

This paper describes the Learning Analytics and Knowledge (LAK) Dataset[4] which represents an unprecedented corpus exposing a near complete collection of bibliographic resources of the particular scientific disciplines 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. Considering the licensing and copyright constraints involved in publishing large amounts of scholarly publications across heterogeneous sources, the LA and EDM discipline lends itself to an ideal use case, as it is a young yet quickly evolving community. Scientific outlets here are still limited to a few main conferences and journals, many of which are open access, allowing for the accumulation of a close to complete corpus spanning all significant publications in the field.

The dataset provides Linked Data about bibliographic metadata as well as full text for all publications. Publication agreements were reached with ACM for publications not already available as open access. The dataset is published and maintained with support of the LinkedUp project[5], the Society for Learning Analytics Research[6] (SoLAR), ACM[7], the L3S Research Center[8] and the Institute for Educational Technology of the National Research Council of Italy[9] (CNR-ITD), with the main goals being (i) facilitating scientific and community analysis of the LA/EDM communities over time and (ii) improving access to scientific literature in said fields, and (iii) providing a general example of open publishing as well as a test-bed for scientometric tools and methods. 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. In addition, the methods and vocabularies used for annotating and exposing the data are describing general practices for publishing bibliographic data beyond mere metadata.

2. Related Work

Publishers of bibliographic data and especially scientific bibliographies have been early adopters of Semantic Web technologies for several years, possibly because of the strong relationship between the fields of library management and information management and the strong use case for sharing scientific publications and related data. That led to a wealth of datasets and vocabularies in the area, where some of the most prominent datasets in the Linked Data cloud today are exposed by organisations such as the British Library (see Linked Open BNB[10]), as well as repositories of research outputs (such as DBLP[11] or the Linked Data Platform from Nature[12]).

That also led to the emergence of vocabularies for bibliographic information, where earlier works include the SwetoDblp ontology [9] and more recent efforts include the BibBase ontology [10], linked also to the Bibliographic Ontology BIBO[13], and the Semantic Web for Research Communities[14] (SWRC) ontologies, two other widely used vocabularies. Schema.org, and the SPAR ontology suite[15] also offer a wide range of concepts and vocabularies in this context, where the WorldCat Linked Data Vocabulary[16] of the OCLC[17] recommends schema.org types.

The Semantic Web Dog Food (SWDF)[18] initiative, using the SWRC vocabulary, aims towards creating a complete Linked Data repository of metadata of papers submitted to conferences associated with the Semantic Web domain. Our endeavour follows a similar approach, collecting publication data from relevant scientific venues (even using same or mapped vocabularies), in the field of Learning Analytics. Different to SWDF or otherwise highly related works as in [10], we aim to enable analyses not only

[4] http://lak.linkededucation.org
[5] http://linkedup-project.eu
[6] http://www.solaresearch.org
[7] http://acm.org
[8] http://www.l3s.de
[9] http://www.itd.cnr.it
[10] http://www.bl.uk/bibliographic/datafree.html#lod
[11] http://datahub.io/dataset/l3s-dblp
[12] http://data.nature.com/
[13] http://bibliontology.com/
[14] http://ontoware.org/swrc/
[15] http://sempublishing.sourceforge.net/
[16] http://oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html
[17] http://oclc.org
[18] http://data.semanticweb.org/

of the metadata, but also of the actual paper content, through also providing access to the full-text of papers, and linking them to other, complementary sources of data.

3. The LAK Dataset - Content, Scope and Maintenance

While we also offer regularly updated dumps (RDF/XML, N-Triples and R[19]), here we specifically discuss the RDF dataset and SPARQL endpoint, accessible as described in Table 1.

Table 1 LAK Dataset facts table

Name	LAK Dataset
Dataset Home	http://lak.linkededucation.org
Schema	http://lak.linkededucation.org/schema/lak.rdf
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
Licence	Creative Commons Attribution (cc-by) for metadata and Open Access graph, special terms for full text of ACM publications (as described in Section 3.1)

3.1. Coverage, data sources, licences

Data, including metadata and full text, is extracted from papers sourced from all editions of the two main conferences in the LA and EDM fields (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 3, 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. Given the variety of sources and data, the data is split into four subgraphs where different license models apply:

1) *http://lak.linkededucation.org/openaccess*: contains the metadata for all open access publications (see type *Open Access* in Table 3).
2) *http://lak.linkededucation.org/openaccess/body*: contains the full text body of all open access publications
3) *http://lak.linkededucation.org/acm*: contains metadata of all ACM publications (see type *ACM* in Table 3)
4) *http://lak.linkededucation.org/acm/body*: contains the full text body of all ACM publications

Further details about the publications in each graph are shown in Table 3. Data from graphs (1) - (2) are available under CC-BY licence[20]. For data in graphs (3) and (4), we have negotiated a formal agreement with ACM to publish, share and enable reuse of the data. We are currently in discussions to decide on a suitable licence and will update the data and respective metadata on the website and our entries in dataset registries such as the DataHub accordingly.

3.2. 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) preprocessing, clean-up and consolidation of the textual information, (3) lifting data into RDF schema (Section 4.1). 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. Further details are provided in 0. At this stage the full text has been extracted without further considering its structure, while ongoing work is concerned with further structuring the text body. Literature references are also extracted and made available in order to support scientometrics based on co-citation networks.

Given the nature of the dataset, new publications are added continuously as these become available, i.e. whenever new proceedings or journal issues of the reflected series are published. Optimisation of the processing pipeline throughout previous years facili-

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

[20] https://creativecommons.org/licenses/by/2.0/

tates a straight-forward and efficient extraction process for new publications.

The ongoing maintenance of the dataset is carried out as a collaborative activity of all partners including the authors of this paper and their institutions, as well as SoLAR, being one of the central organisations driving the advancement of the LA discipline. 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.

4. Schema, Mappings and Interlinking

4.1. Schema

For each publication the following features are extracted: title, authors, keywords, abstract, text body, references, publication venue (journal/conference proceedings). To ensure wide interoperability of the data, we have adapted Linked Data best practices[21] and investigated widely used vocabularies for the annotation of involved concepts as discussed in Section 2. Preliminary work in 0 investigated most frequent schemas, particularly for educational datasets, and additionally Linked Data vocabulary usage statistics[22] have been investigated. While the scope of our data model is not covered by a single vocabulary alone, we have opted for using established vocabularies for each specific type and predicate and included mappings between the chosen vocabularies as well as other overlapping ones. The schema is accessible at http://lak.linkededucation.org/schema/lak.rdf [23].

The majority of schema elements are based on BIBO, FOAF[24], SWRC, Schema.org, as reported in Table 2. 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. Choice of vocabulary terms was influenced by the Web-wide adoption and maturity of the used schemas and their overlap with our data model. The combination of terms led to the emergence of new type and predicate mappings, which have been represented as explicit mappings using the predicates *owl:equivalentClass* and *owl:equivalentProperty* together with type and property inheritance statements. Mappings rely largely on established recommendations from the vocabulary owners, such as BIBO/schema.org mappings recommended by schema.org[25].

Table 2. Schemas and namespaces used in LAK Dataset

Vocab.	Namespace URL
foaf	http://xmlns.com/foaf/0.1/
swrc	http://swrc.ontoware.org/ontology#
schema.org	http://schema.org/
bibo	http://purl.org/ontology/bibo/
swc	http://data.semanticweb.org/ns/swc/ontology#
dc	http://purl.org/dc/elements/1.1/
dcterms	http://purl.org/dc/terms/

The main classes and predicates are listed in Figure 1, and Tables 4 and 5. By relying entirely on established and frequently used types and properties, we aim for a high reusability of the data.

[21] http://www.w3.org/TR/ld-bp/#VOCABULARIES

[22] http://stats.lod2.eu/stats

[23] While this URL always refers to the latest version of the schema, current and previous versions are also accessible, for instance, via http://lak.linkededucation.org/schema/lak-v0.2.rdf

[24] http://xmlns.com/foaf/spec/

[25] http://schema.rdfs.org/mappings.html

Table 3 Academic publications in the dataset

Publication Venue	# Papers	Type	Named Graph URI
Proceedings of the ACM International Conference on Learning Analytics and Knowledge (LAK) (2011-2014)	166	ACM	http://lak.linkededucation.org/acm http://lak.linkededucation.org/acm/body
Proceedings of the International Conference on Educational Data Mining (2008-2014)	463	Open Access	http://lak.linkededucation.org/openaccess http://lak.linkededucation.org/openaccess/body
Special issue on "Learning and Knowledge Analytics": Educational Technology & Society, edited by George Siemens & Dragan Gašević), 2012, 15, (3), 1-163.	10	Open Access	
Journal of Educational Data Mining (2009-2014)	29	Open Access	
Journal of Learning Analytics (2014)	16	Open Access	

Mappings were evaluated for consistency (using the HermiT reasoner[26]) with the involved schemas.

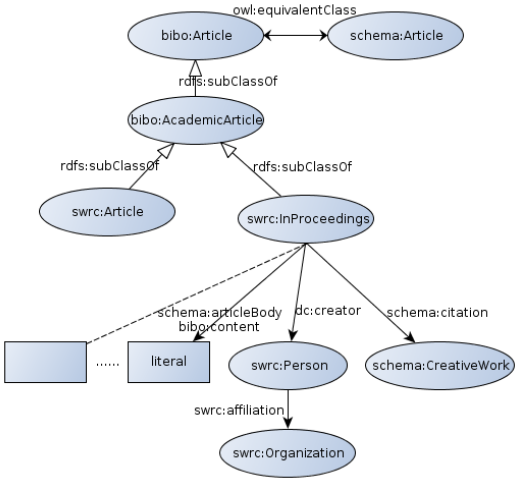

Fig. 1: Key classes and properties used in the LAK Dataset (conference proceedings only)

The following Table 4 provides a general overview of the number of represented entities per type in the LAK dataset.

Table 4. Entity population in the LAK Dataset

Concept	**ofType**	#
Reference	schema:CreativeWork	7885
Author	swrc:Person	1214
Conference Paper	swrc:InProceedings	697
Organization	swrc:Organization	365
Journal Paper	swrc:Article	45
Conference Proceedings	swrc:Proceedings	15
Journal Issue	bibo:Issue	9
Journal	bibo:Journal	2

Table 5 summarizes the most frequently populated properties.

Table 5 Most frequently populated properties in the LAK Dataset

Domain	**Property**	**Range**	**#**
schema:Article	schema:citation	schema:CreativeWork	10828
swrc:InProceedings	dc:subject	literal	3392
foaf:Agent	foaf:made	swrc:InProceedings	2199
foaf:Person	rdfs:label	literal	1583
foaf:Agent	foaf:sha1sum	literal	1341
swrc:Person	swrc:affiliation	swrc:Organization	1293
foaf:Person	foaf:based_near	geo:SpatialThing	1243
schema:Article	schema:articleBody	literal	698
bibo:Article	bibo:abstract	literal	697
bibo:Issue	bibo:hasPart	bibo:Article	45
swrc:Proceedings	swc:relatedToEvent	swc:ConferenceEvent	14
bibo:Journal	bibo:hasPart	bibo:Issue	9

[26] http://hermit-reasoner.com

4.2. Inter-Dataset Links

While bibliographic metadata 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 that LAK is considered a sub-discipline of Computer Science (CS), we have particularly considered the datasets DBLP and Semantic Web Dog Food. While DBLP allows us to link authors to their corresponding representation in a more exhaustive bibliographic CS 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 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 broad 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 Computer Science literature as a whole. For instance, in DBLP and Semantic Web Dogfood, LAK publications are not exhaustively represented, references and full text are 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. Such links enable 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 (Name Entity Recognition) methods (DBpedia Spotlight), allowing to link for instance keywords such as "educational gaming" to corresponding DBpedia entities, such as *http://dbpedia.org/resource/Educational_game*, an example taken from a particular EDM2014 paper[27].

The following Figure 2 depicts the links of resolved or enriched LAK entities.

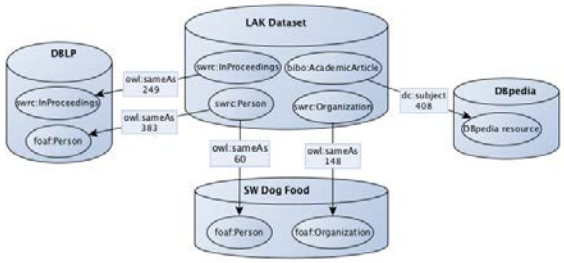

Fig. 2 : Links in the LAK Dataset[28]

With respect to inlinks, the dataset is referenced by the LinkedUp catalog[29] and the majority of its resources are referenced by the Linked Dataset Profiles[30] dataset, further described in [5]. Additional inlinks might have been generated by the works described in [7][8].

5. Query and exploration

Some example queries[31] which demonstrate the datasets usefulness with respect to the reported objectives (Section 1) are shown 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).

```
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX dbpedia:<http://dbpedia.org/resource/>

select distinct ?paper ?subject
from <http://lak.linkededucation.org/openaccess>
from <http://lak.linkededucation.org/acm>
where {
 ?paper dc:subject dbpedia:Educational_game .
 ?paper dc:subject ?subject .
}
```

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 {
graph <http://lak.linkededucation.org/acm>
{ <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 3 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
from <http://lak.linkededucation.org/openaccess>
from <http://lak.linkededucation.org/acm>
where {
 ?paper a swrc:InProceedings .
 ?paper npg:hasCitation ?reference
} order by desc(?count) limit 10
```

Listing 3: Retrieving influential publications by means of the most cited papers.

[27] http://data.linkededucation.org/resource/lak/conference/edm2014/paper/580

[28] A high resolution version of this figure is available at: http://lak.linkededucation.org/lak/lak_links.png

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

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

[31] Additional queries available at: http://lak.linkededucation.org/?page_id=351

6. Applications, impact & usage

The LAK Dataset has received considerable attention and support from organisations such as SoLAR, which also advertises the dataset for its own purposes[32]. Throughout the last years, the dataset has emerged into a central resource for researchers in the LA and EDM field and beyond, documented by a variety of research publications which make use of the data. Including the proceedings [7][8], the authors already are aware of 16 scientific publications[33] which make use of the LAK dataset. While the value of the data for the LA and EDM fields is obvious, the dataset also provides an unprecedented resource for general investigations into scientometrics and in particular, their evolution over time, given the almost complete coverage of the entire research corpus of the covered communities.

The dataset also forms the basis of the LAK Data Challenge, organised by the authors and a team of researchers affiliated with SoLAR, LinkedUp[34] and associated organisations as an annual competition (now in its third year). It is co-located with the ACM LAK conferences (LAK2013[35], LAK2014[36]) with currently open calls for the 2015 edition, directly supported by the steering board of the LAK conference. 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.

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.

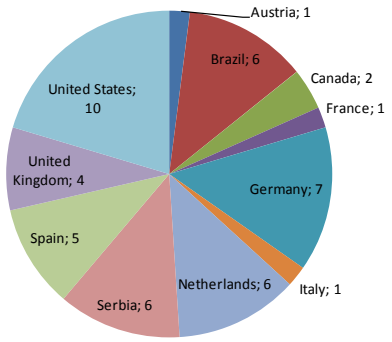

Fig. 3. LAK Data Challenge submissions - authors per country

Applications are further described in the proceedings of the 2013 and 2014 LAK Data Challenge editions [7][8] and are available online[37,38]. 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 in *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" [3], or near complete scientometric environments such as DEKDIV [4] (depicted in Fig. 4).

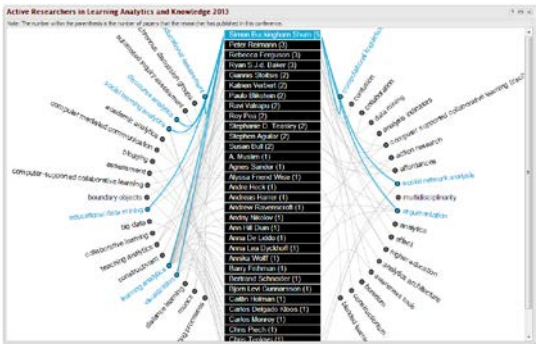

Fig. 4. DEKDIV active LA researchers exploration

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

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

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

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

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

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

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

The latter combines 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, where Fig. 4 shows the most frequent authors with regards to a specific set of topics.

Next to these applications, the dataset and some of its applications have been endorsed and supported by SoLAR and ACM, where current discussions are geared towards embedding some of the described applications into their more general libraries and platforms. In addition, as joint activity of the authors and SoLAR, current work aims at expanding the dataset with actual learning analytics research data, i.e. data usually used in the captured publications. The joint vision is to provide a near-complete corpus which provides not just the actual scientific publications in structured formats, but also to a larger extent, their used raw research datasets. This is meant to further facilitate LA & EDM research and open access to research publications and data in general.

7. Discussion & Future Work

In this paper, we have presented (a) the LAK Dataset, as a particular resource which enables the exemplary investigation and analysis of the evolution of scientific disciplines and the validation of scientometric methods and tools, and (b) a vocabulary, collection of mappings and linking practices for adoption in similar efforts, towards a wider movement engaging in the publication of open and machine-processable scholarly resources.

While, according to the 5-star classification[39] of LOD and Vocabulary use (see also [3]) the LAK Dataset qualifies as a 5-star dataset, there are known shortcomings which the authors are addressing as part of ongoing and future work. The extraction process is not entirely flawless and, depending on the quality of the source PDFs, had in some cases required manual adjustment. Given that the automated co-reference resolution had to consider particular drawbacks, we specifically preferred high precision in favor of recall, to ensure a knowledge graph which is as correct as possible, rather than as complete as possible. We are currently looking into more sophisticated entity interlinking methods, in order to further increase the linking to related entities in other datasets. In addition, the extraction of references and full text is so far in a preliminary stage, providing both references and text body in a fairly 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. Similarly, we are working on providing a more detailed structuring of the text body, applying the Document Components Ontology (DoCO)[40] in order to distinct different textual components, such as headings, captions or sections.

Additional insights were gained from the vocabulary definition process. Given the specific scope of our dataset, covering bibliographic metadata and full text, it has been necessary to combine elements from different, partially overlapping vocabularies. We relied on established vocabularies to represent the different involved notions. Due to cross-vocabulary statements, implicit type and predicate mappings emerged which were explicitly represented through dedicated mapping statements. Next to these, additional mappings were introduced to ensure wide interoperability of the data. Given the complex relationships emerging from such vocabulary usage, assessing the compliance of new introduced cross-vocabulary mappings is crucial to eliminate any conflicts. In particular the evolution of external vocabularies might pose issues, where continuous monitoring is required to ensure compliance at all times. To this end, the encapsulation of all schema-level statements in our datasets is meant to serve as a starting point for similar efforts, for instance, for exposing bibliographic data in other disciplines.

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 [7][8]. Given that the actual text body of publications contains substantial information but is yet still missing from the majority bibliographic Linked Data, we would like to encourage work on similar efforts, i.e. the creation of bibliographic datasets containing both metadata and the actual content. In this context, our work provides a set of practices for related efforts in other scientific areas. This would allow a more direct processing and analysis of scientific works across disciplines. Furthermore, applying such approaches 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,

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

[40] http://purl.org/spar/doco

a topic widely discussed not only in the Semantic Web community but also supported by corresponding European directives.

References

[1] D. Taibi, S. Dietze, Fostering Analytics on Learning Analytics Research: the LAK Dataset, in: Eds: M. d'Aquin, S. Dietze, H. Drachsler, E. Herder, D. Taibi, CEUR WS Proceedings Vol. 974, Proceedings of the LAK Data Challenge, held at LAK2013 - 3rd International Conference on Learning Analytics and Knowledge, Leuven, Belgium, 2013.

[2] M. d'Aquin, A. Adamou, S. Dietze, Assessing the Educational Linked Data Landscape, Proceedings of ACM Web Science 2013 (WebSci2013), ISBN: 978-1-4503-1889-1, Paris, France, ACM, 2013, 43-46.

[3] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, C. Vardeman II, Five Stars of Linked Data Vocabulary Use, Semantic Web, 5 (3), 2014, 173-176.

[4] B. P. Nunes, B. Fetahu, S. Dietze, M. A. Casanova, Cite4Me: A Semantic Search and Retrieval Web Application for Scientific Publications, in: Eds. E. Blomqvist, T. Groza, Proceedings of the ISWC 2013 Posters & Demonstrations Track, a track within the 12th International Semantic Web Conference, CEUR WS Proceedings Vol. 1035, Sydney, Australia, 2013.

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

[6] B. Fetahu, S. Dietze, B.P. Nunes, M.A. Casanova, D. Taibi, W. Nejdl, A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles, in: The Semantic Web: Trends and Challenges, Proceedings of the 11th Extended Semantic Web Conference (ESWC2014), Eds.: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A., Lecture Notes in Computer Science, Vol. 8465, 519-534, Springer International Publishing (2014).

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

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

[9] B. Aleman-Meza, F. Hakimpour, I.B. Arpinar, A.P. Sheth, SwetoDblp Ontology of Computer Science Publications, Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 5(3), 2007, 151-155.

[10] R. S. Xin, O. Hassanzadeh, C. Fritz, S. Sohrabi, R. J. Miller, Publishing Bibliographic Data on the Semantic Web using BibBase, Semantic Web Journal, IOS Press, Amsterdam, The Netherlands, 4(1), 2013, 15-22.