Chapter 29: Linked Data for Learning Analytics: The Case of the LAK Dataset

Davide Taibi[1], Stefan Dietze[2]

[1] Istituto per le Tecnologie Didattiche, Consiglio Nazionale delle Ricerche, Italy
[2] L3S Research Center, Germany

DOI: 10.18608/hla17.029

ABSTRACT

The opportunities of learning analytics (LA) are strongly constrained by the availability and quality of appropriate data. While the interpretation of data is one of the key requirements for analyzing it, sharing and reusing data are also crucial factors for validating LA techniques and methods at scale and in a variety of contexts. Linked data (LD) principles and techniques, based on established W3C standards (e.g., RDF, SPARQL), offer an approach for facilitating both interpretability and reusability of data on the Web and as such, are a fundamental ingredient in the widespread adoption of LA in industry and academia. In this chapter, we provide an overview of the opportunities of LD in LA and educational data mining (EDM) and introduce the specific example of LD applied to the Learning Analytics and Knowledge (LAK) Dataset. The LAK dataset provides access to a near-complete corpus of scholarly works in the LA field, exposed through rigorously applying LD principles. As such, it provides a focal point for investigating central results, methods, tools, and theories of the LA community and their evolution over time.

Keywords: Linked data, LAK dataset

Learning analytics (LA)[1] and educational data mining (EDM)[2] have gained increasing popularity in recent years. However, large-scale take-up in educational practice as well as significant progress in LA research is strongly dependent on the quantity and quality of available data. Being able to interpret and understand data about learning activities, including respective knowledge about the learning domain, subjects, or skills is a prerequisite for carrying out higher-level analytics. However, data as generated through learning environments is often ambiguous and highly specific to a particular learning scenario and use case, often using proprietary terminologies or identifiers for both understanding and interpreting learning-related data within the scenario, and even more, across organizational and application boundaries.

LD principles (Bizer, Heath, & Bernes-Lee, 2009) have emerged as a de facto standard for exposing data on the Web and have the potential to improve both the quantity and quality of LA data substantially by 1) enabling *interpretation* of data and 2) Web-wide *sharing* of datasets across scenarios and institutional boundaries. Facilitated through established W3C standards such as RDF and SPARQL, LD has gained significant popularity throughout the last decade, with over 1000 datasets in the recent Linked Open Data Crawl[3] alone. LD and its offspring see widespread adoption through all sorts of entity-centric approaches, such as the use of knowledge graphs for facilitating Web search, a common practice in major search engines such as Google or Bing, or the increasing adoption of Microdata and RDFa[4] for annotating Web pages with structured facts. This also led to the emergence of a growing Web of educational data (d'Aquin, Adamou, & Dietze, 2013), substantially facilitated by the availability of shared vocabularies for educational purposes and knowledge graphs such as DBpedia[5] or Freebase[6] for enriching and disambiguating data.

We argue that LD principles can act as a fundamental facilitator for scaling up LA research (d'Aquin, Dietze,

[1] http://solaresearch.org
[2] http://www.educationaldatamining.org
[3] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/
[4] https://www.w3.org/TR/xhtml-rdfa-primer/
[5] http://dbpedia.org
[6] http://freebase.org

Herder, Hendrik, & Taibi, 2014), as well as improving performance of LA tools and methods by enabling 1) the non-ambiguous interpretation of learning data (d'Aquin & Jay, 2013) and 2) the widespread sharing of the data used for evaluating and assessing LA methods and tools in research and educational practice. After a brief summary of LD use in education, we will introduce the successful application of LD principles in the LA context of the LAK dataset.[7] The LAK dataset represents, on the one hand, a representative example of successfully applying LD principles to facilitate research in LA and, on the other, constitutes an important resource in its own right by providing access to a near-complete corpus of LA and EDM research. This is followed by a set of examples that demonstrate the benefits of applying LD principles by showcasing how new insights can be generated from such a corpus and, at the same time, provide insights into observable trends and topics in LA and EDM.

LINKED DATA IN LEARNING AND EDUCATION

Distance teaching and openly available educational data on the Web are becoming common practices with public higher education institutions as well as private training organizations realizing the benefits of online resources. This includes data 1) about *learning resources*, ranging from dedicated educational resources to more informal knowledge resources and content, and 2) data about *learning activities*.

LD principles (Heath & Bizer, 2011) offer significant opportunities for sharing, interpreting, or enriching data about both resources and activities in learning scenarios. Essentially, LD principles rely on a common graph-based representation format, the so-called Resource Description Framework (RDF),[8] a common query language (SPARQL[9]) and most notably, the use of dereferenceable URIs to name things (i.e., entities). This last feature is a key facilitator for LD as it enables the unique identification of any entity in any dataset across the Web, and hence links data across different datasets. This facilitates, for instance, an entity representing LA in the DBpedia dataset[10] being linked with co-references in non-English DBpedias[11] or co-references in other datasets such as Freebase.[12]

These principles have enabled the emergence of a global graph of LD on the Web, including cross-domain data such as DBpedia, WordNet RDF,[13] or the data.gov.uk initiative, as well as domain-specific expert vocabularies, for instance, of data about cultural heritage (e.g., the Europeana dataset[14]). This has also led to the creation of an embryonic "Web of Educational Data" (see d'Aquin et al., 2013; Taibi, Fetahu, & Dietze, 2013; and Dietze et al., 2013, for an overview) including data from institutions such as the Open University (UK)[15] or the National Research Council (Italy),[16] as well as publicly available educational resources, such as the mEducator – Linked Educational Resources (Dietze, Taibi, Yu, & Dovrolis, 2015). Initiatives such as LinkedEducation.org,[17] LinkedUniversities.org,[18] and LinkedUp[19] have provided first efforts to bring together people and works in this area. In this context, the LinkedUp Catalog[20] is an unprecedented collection of publicly available LD relevant to educational scenarios, containing data about dedicated open educational resources (OER), such as Open Courseware (OCW) or mEducator datasets, data about bibliographic resources, or metadata about other knowledge resources.

While data about learning activities is not frequently available and data sharing even less so, LD has been adopted to facilitate representation of social and activity or attention data (Dietze, Drachsler, & Giordano, 2014); Ben Ellefi, Bellahsene, Dietze, and Todorov (2016) provide a thorough overview. In the field of LA, LD principles can substantially improve the disambiguation, interpretation, and understanding of data (as documented by d'Aquin & Jay, 2013; d'Aquin et al., 2014). Reference knowledge graphs, domain-specific or cross-domain, can significantly improve the interpretation and analytical processes of captured learning analytics data by disambiguating and enriching data, for instance, about subjects or competencies. This can improve the performance of learning analytics methods and tools within specific scenarios (d'Aquin & Jay, 2013).

Certain limitations are apparent, however, when dealing with reasoning-based approaches such as Semantic Web technologies. Given the computational demands of interpreting and reasoning on knowledge representations, LD-based approaches are known to be less scalable than traditional RDBMS-based[21] methods. However, given the maturity of existing RDF storage and reasoning engines, this applies specifically to very large-scale datasets, which are less frequent in LA and EDM settings. Other issues include the lack of links, the misuse of schema terms, or the lack of semantic and syntactic quality of exposed data. However, these issues are by no means exclusive or specific to LD-based datasets but prevail across data management

[7] http://lak.linkededucation.org
[8] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
[9] http://www.w3.org/TR/rdf-sparql-query/
[10] http://dbpedia.org/page/Learning_analytics
[11] http://fr.dbpedia.org/resource/Analyse_de_l'apprentissage
[12] http://rdf.freebase.com/ns/m.0crfzwn
[13] http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/
[14] http://ckan.net/package/europeana-lod
[15] http://data.open.ac.uk
[16] http://data.cnr.it
[17] http://linkededucation.org
[18] http://linkeduniversities.org
[19] http://linkedup-project.eu
[20] http://data.linkededucation.org/linkedup/catalog/
[21] Relational database management system.

Table 29.1. Edited Excerpt of Discourse Data Coded in ENA Format

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 (EDM) (2008–2014)	463	Open Access	http://lak.linkededucation.org/openaccess http://lak.linkededucation.org/openaccess/body
Special 2012 issue on "Learning and Knowledge Analytics" edited by George Siemens & Dragan Gašević: Educational Technology & Society, 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	
Proceedings of the LAK data Challenge (2013–2014)	13	Open Access	

technologies of all kinds.

Hence, sharing LA data according to LD principles has the potential to boost the adoption and improvement of LA tools and methods significantly by enabling their evaluation across a range of real-world datasets and scenarios.

THE LAK DATASET: A LINKED DATA CORPUS FOR THE LEARNING ANALYTICS COMMUNITY

In order to provide a best-practice example of adopting LD principles for sharing LA data, we introduce the LAK dataset, a joint effort of an international consortium consisting of the Society for Learning Analytics Research (SoLAR), ACM, the LinkedUp project, and the Educational Technology Institute of the National Research Council of Italy (CNR-ITD). The LAK dataset constitutes a near complete corpus of collected research works in the areas of LA and EDM since 2011, where LD principles have been applied to expose both metadata and full texts of articles (Dietze, Taibi, & d'Aquin, 2017). As such, the corpus enables unprecedented research on the scope and evolution of the LA community. Here, Table 29.1 reports an overview of the publications included in the LAK dataset. Given the variety of sources, the data is split into four subgraphs (last column of Table 29.1 where different license models apply).[22]

To ensure wide interoperability of the data, we have adapted LD best practices[23] and investigated widely used vocabularies for the representation of scientific publications. The scope of our data model is not covered by a single vocabulary alone. For this reason, we opted for using established vocabularies such as BIBO, FOAF,[24] SWRC, and Schema.org for all represented terms and included mappings between the chosen vocabularies as well as other overlapping ones. [25] The choice of vocabulary terms was influenced by the Web-wide adoption and maturity of the used schemas and their overlap with our data model. Table 29.2 reports the concepts represented in the LAK dataset and their population while Table 29.3 summarizes the most frequently populated properties.

Table 29.2. Entity Population in the LAK Dataset

Concept	Type	#
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

Exploiting inherent features of LD, the LAK dataset is enriched with entity links to other datasets, for instance to provide links to author and publication venue co-references and complementary information. In particular, links with the Semantic Web Dog Food (SWDF)[26] dataset and DBLP provide additional information about authors and venues in the LAK dataset,

[22] Data from graphs http://lak.linkededucation.org/openaccess/* are available under CC-BY licence. For data in graphshttp://lak.linkededucation.org/acm/*, we have negotiated a formal agreement with ACM to publish, share, and enable reuse of the data for research purposes. https://creativecommons.org/licenses/by/2.0/

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

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

[25] The currently implemented schema is available at http://lak.linkededucation.org/schema/lak.rdf 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

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

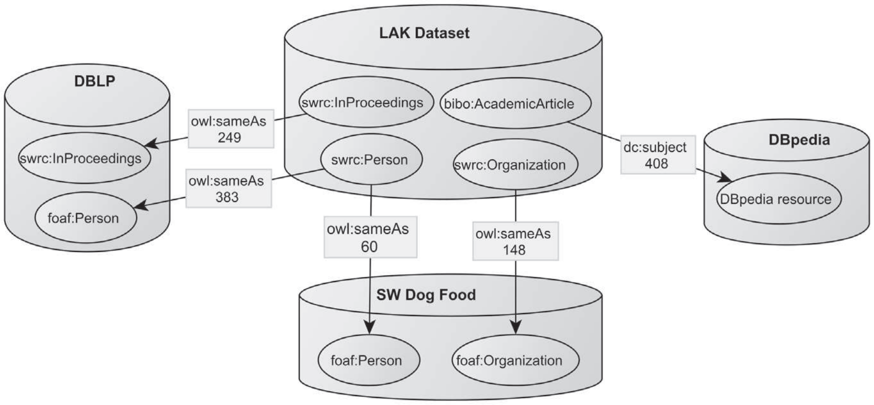

Figure 29.1. Interlinking the LAK dataset.

such as their wider scientific activity and impact. This is useful, for instance, to complement the highly focused nature of the LAK dataset, which by definition has a narrow scope (LA and EDM) and would otherwise limit research to activities within that very community. On the other hand, such established links complement existing corpora with data contained in the LAK dataset by 1) enriching the limited metadata with additional properties and 2) containing additional publications not reflected in DBLP or the Semantic Web Dog Food, creating a more comprehensive knowledge graph of Computer Science literature as a whole.

Table 29.3. Entity Population 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

Additional outlinks were created to DBpedia as reference vocabulary. To allow for a more structured retrieval and clustering of publications according to their topic-wise similarity, we have linked *keywords*, provided by authors, to their corresponding entities in DBpedia, thereby using DBpedia as reference vocabulary for paper topic annotations. Figure 29.1 depicts the links of resolved or enriched LAK entities.

Given the nature of LD, establishing such links has been merely a matter of looking up LAK dataset entities and adding *owl:sameAs* statements, which refer to the IRIs of co-referring entities in DBLP, SW Dog Food, and DBpedia. Hence, this process is enabled by fundamental principles of LD, such as using URIs to identify things and using SPARQL queries to demonstrate the key motivation: the creation of a global data graph rather than isolated datasets.

LINKED DATA-ENABLED INSIGHTS INTO THE LAK CORPUS: SCOPE AND TRENDS OF LEARNING ANALYTICS RESEARCH

To illustrate the exploitation of LD principles implemented in the LAK dataset, we introduce some simple analysis enabled through the inherent links within the dataset, as described above. While a wide range of additional investigations can be found in the applications and publications of the LAK Data Challenge,[27] here we focus on a set of very simple investigations and research questions. These can be answered merely by combining SPARQL queries on the LAK dataset and interlinked datasets, and by exploiting the links between co-references described in the earlier section. These analyses are primarily aimed at demonstrating the ease of answering complex research questions by combining data from different sources. In particular, we investigate questions related to the following:

1. The research background and focus of researchers in the LA community, in order to shape a picture of the constituting disciplines and areas of this comparably new research area: This investigation

[27] See the application and publication sections at http://lak.linkededucation.org

Table 29.4. Top 20 Conferences and Journals in which LAK 2011 Conference Authors Previously Published

Conference or Journal	DBLP resource	%
Intelligent Tutor Systems Conference	http://dblp.l3s.de/d2r/resource/conferences/its	24.49
Educational Data Mining Conference	http://dblp.l3s.de/d2r/resource/conferences/edm	12.05
Artificial Intelligence in Education Conference	http://dblp.l3s.de/d2r/resource/conferences/aied	11.15
European Conference on Technology Enhanced Learning	http://dblp.l3s.de/d2r/resource/conferences/ectel	7.31
International Conference on Advanced Learning Technologies	http://dblp.l3s.de/d2r/resource/conferences/icalt	5.51
AAAI Conference on Artificial Intelligence	http://dblp.l3s.de/d2r/resource/conferences/aaai	4.36
UM conference	http://dblp.l3s.de/d2r/resource/conferences/um	4.36
IEEE International Conference on Data Mining	http://dblp.l3s.de/d2r/resource/conferences/icdm	3.72
Conference on Knowledge Discovery and Data Mining	http://dblp.l3s.de/d2r/resource/conferences/kdd	3.59
International Journal of Artificial Intelligence in Education	http://dblp.l3s.de/d2r/resource/journals/aiedu	2.56
ETS journals	http://dblp.l3s.de/d2r/resource/journals/ets	2.31
Conference on Computer Supported Collaborative Learning	http://dblp.l3s.de/d2r/resource/conferences/cscl	2.31
International Conference on Machine Learning	http://dblp.l3s.de/d2r/resource/conferences/icml	2.31
Journal of Universal Computer Science	http://dblp.l3s.de/d2r/resource/journals/jucs	2.18
International Conference of the Learning Sciences	http://dblp.l3s.de/d2r/resource/conferences/icls	2.18
ACM CHI Conference	http://dblp.l3s.de/d2r/resource/conferences/chi	2.18
World Wide Web conference	http://dblp.l3s.de/d2r/resource/conferences/www	2.18
AH conference	http://dblp.l3s.de/d2r/resource/conferences/ah	1.92
International Joint Conference on Artificial Intelligence	http://dblp.l3s.de/d2r/resource/conferences/ijcai	1.67
Machine Learning Journal	http://dblp.l3s.de/d2r/resource/journals/ml	1.67

exploits information about LA researchers' publication activity in other areas by using DBLP data.

2. The importance of key topics in the LA field and their evolution over time: This investigation exploits the topic (or category) mapping of LA keywords (or entities) in DBpedia and their relationships.
3. The apparent links between the LA and LD communities that can be derived from the data.

In both cases, our analysis has been conducted by taking into account all the publications of the LAK conferences from 2011 to 2014, available in the LAK dataset, in order to study the evolution of LA research over the years.

Who Makes Up the Learning Analytics Community?Publication Activities of LA Researchers

The development of LA has been influenced by the intersection of numerous academic disciplines such as machine learning, artificial intelligence, education technology, and pedagogy (Dawson, Gašević, Siemens, & Joksimović, 2014). For this reason, since its first edition, the LAK conference has drawn the attention of researchers from different scientific fields, each contributing their definitions, terminologies, and research methods, and thereby shaping the definition of what LA is. The core data of the LAK dataset, being limited to LA-related publication activities exclusively, does not enable any analysis into the origin and research background of contributing researchers. The LD nature of the corpus, however, provides meaningful connections that can be exploited to infer such new knowledge. In fact, by linking the resources representing the authors in the LAK dataset with the authors in the DBLP dataset, it is feasible with a few SPARQL queries to extract further information about the fields of interest of the authors.[28]

For all LAK authors in each year from 2011 to 2014, we analyzed the number of publications in previous conferences and journals by first 1) obtaining all authors of a respective year in the LAK dataset and 2) retrieving their previous publication venues (journals,

[28] The 36% of the 1214 authors in the LAK dataset is linked with the correspondent resource in the DBLP (86%) and SWDF datasets (14%).

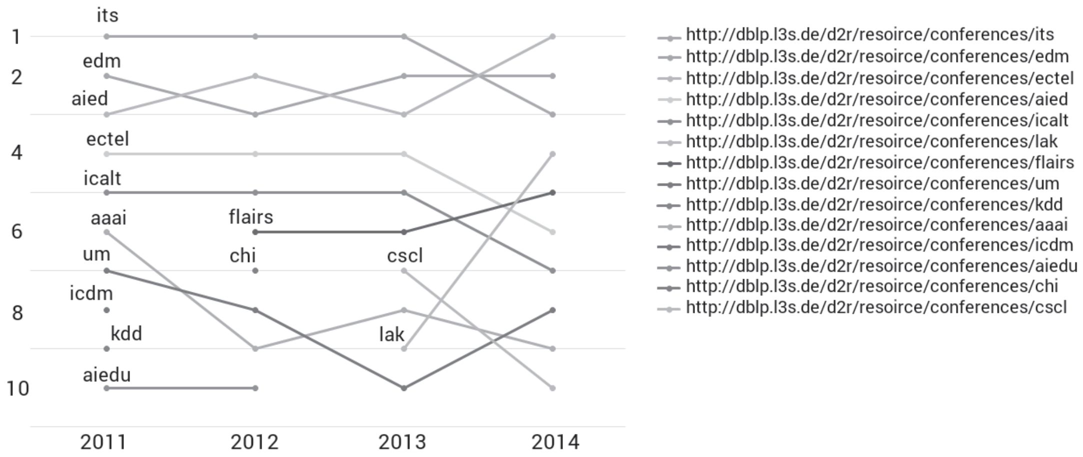

Figure 29.2. Interlinking the LAK dataset.

conferences) from DBLP. The top 20 conferences and journals for 2011 are reported in Table 29.4. This table highlights that, in its first edition, the LAK conference has mainly involved authors with previous publications related to the *Intelligent Tutor Systems*, *Educational Data Mining*, *Artificial Intelligence*, and *Technology Enhanced Learning* conferences. From a technical point of view, interlinks between the LAK and DBLP datasets were created as follows: LAK authors are linked with their co-references in the DBLP dataset through the *owl:sameAs* property The DBLP authors in turn are connected with their publications in previous conferences and journals respectively through the *swrc:series* and the *swrc:journal* properties. The execution of a federated query involving the two datasets allows us to deduce information about the number of publications of LAK authors in previous conferences and journals.

In Figure 29.2, we report the rank of the top 10 conferences and journals in which LAK authors have published from 2011 to 2014. The top three positions are clearly occupied by the ITS (Intelligent Tutor Systems), EDM (Educational Data Mining), and AIED (Artificial Intelligence in Education) conferences/journals. Starting in 2013, the LAK conference appears in the top 10, growing in importance in 2014, indicating the constitution of a significant community in its own right. That same year saw an increasing number of papers published in the FLAIRS (Florida Artificial Intelligence Research Society) conference proceedings. Publications from EC-TEL (European Conference on Technology Enhanced Learning) and ICALT (International Conference on Advanced Learning Technologies) conferences also appear at the top of the list, with a slight inflection in the last year.

Which Topics Make Up the LA Field? How Does Topic Distribution Change Over Time?

In contrast to the previous investigations, interlinking LAK conference publications with relevant DBpedia

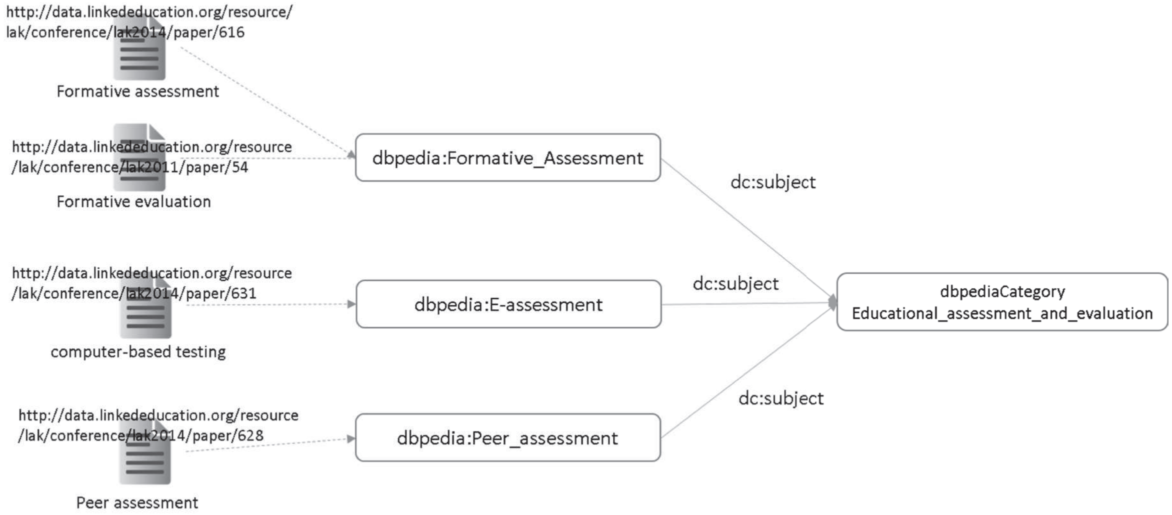
http://data.linkededucation.org/resource/
lak/conference/lak2014/paper/616
Formative assessment
http://data.linkededucation.org/resource
/lak/conference/lak2011/paper/54
Formative evaluation
http://data.linkededucation.org/resource
/lak/conference/lak2014/paper/631
computer-based testing
http://data.linkededucation.org/resource
/lak/conference/lak2014/paper/628
Peer assessment
dbpedia:Formative_Assessment
dbpedia:E-assessment
dbpedia:Peer_assessment
dc:subject
dc:subject
dc:subject
dbpediaCategory
Educational_assessment_and_evaluation

Figure 29.3. Interlinking publications and DBPedia.

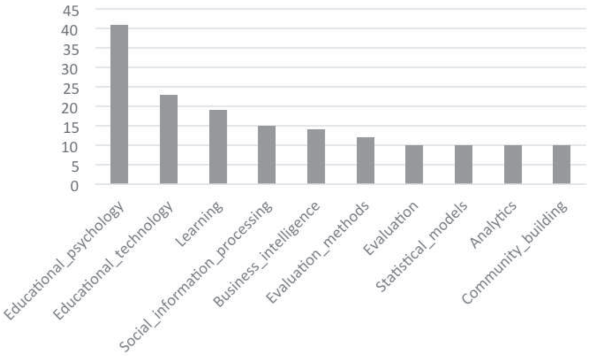

Figure 29.4. Top-10 DBPedia categories for the LAK dataset (publications from 2011-2014).

entities allows us to investigate the semantics of topics covered by analyzing the DBpedia knowledge graph and the inherent links of entities and categories. This, for instance, enables us to identify the overlap of LAK papers with other disciplines, such as *Computer Science*, *Statistics*, *Technology Enhanced Learning*, or *Data Analysis*.

As described in the previous section, links between the LAK dataset and DBpedia entities were established by disambiguating terms (keywords) through state-of-the-art NER (Name Entity Recognition) methods (DBpedia Spotlight). This allowed us to link keywords – for instance, "computer-based testing" and "formative evaluation" – respectively to the corresponding DBpedia entities, *http://dbpedia.org/resource/E-assessment* and *http://dbpedia.org/resource/Formative_assessment*[29] (see Figure 29.3). Each DBpedia entity, in turn, is connected through the dc:subject property, to its corresponding DBpedia categories; for instance, the category Educational_assessment_and_evaluation is the dc:subject of DBpedia resources: Formative_assessment, E-assessment, Peer_assessment, and Educational_evaluation, just to name a few. In this way, papers can be clustered according to their structural similarity within the DBpedia graph. The list of top 10 DBpedia categories with the highest frequency value in the LAK dataset is shown in Figure 29.4.[30]

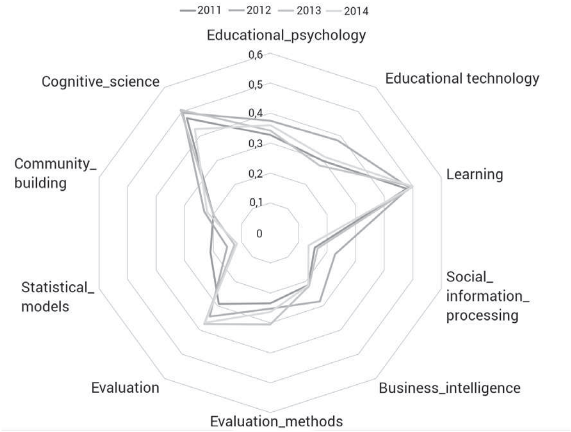

Figure 29.5. Evolution of the top-10 categories over time.

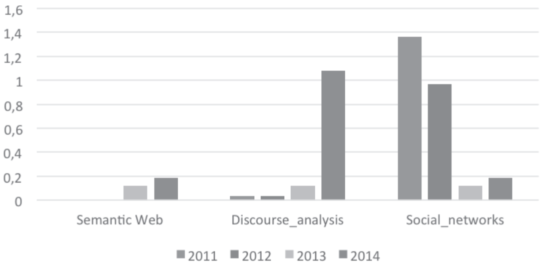

Figure 29.6. Evolution of selected categories, 2011–2014.

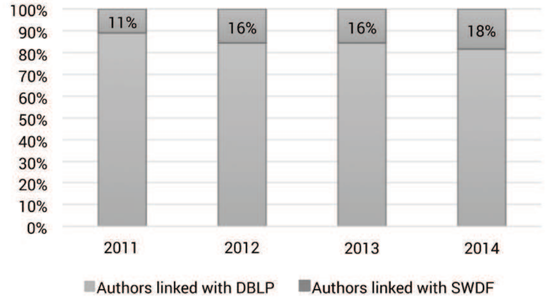

Figure 29.7. Percentage of LAK authors represented within DBLP and Semantic Web Dog Food.

Starting from this set of top-10 most frequent categories over 2011 to 2014, we evaluated the distances between all the DBpedia categories extracted for each conference year and the categories included in this set of "base categories." The SKOS[31] properties used by the DBpedia category graph to represent relationships between categories were exploited to compute this distance. For example, the distance between *E-Learning* and *Educational_technology* is 2, since *Educational_technology* is *skos:broader* of *Distance_education* and, in turn, *Distance_education* is *skos:broader* of *E-learning*.

The relation between DBpedia categories and LAK conference papers also makes it easier to trace the trend of topics covered by LAK publications over the years. The radar chart in Figure 29.5 provides an overview of the calculated average distance between *all categories extracted for each conference year and each category included in the "base category" set*. From the analysis of the figure, the following considerations arise:

- *Educational_technology* played a key role in 2012 but in other years more specialized categories

[29] See the following papers: http://data.linkededucation.org/resource/lak/conference/lak2011/paper/54 and http://data.linkededucation.org/resource/lak/conference/lak2014/paper/616
[30] References to DBpedia categories are in the form: http://dbpedia.org/page/Category:Educational_psychology, http://dbpedia.org/page/Category:Educational_technology and so on.
[31] http://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html

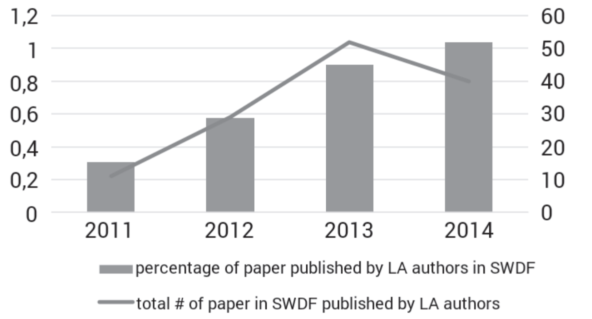

Figure 29.8. Trend of previous publications of LAK authors in SW conferences.

gained importance

- The fairly broad categories of *Learning* and *Evaluation* have had the greatest relevance in all years
- The relevance of *Evaluation* has an increasing trend over the years, with a sensitive increment from 2011 to 2012.
- *Statistical_models* peaked in 2011 and decreased in subsequent years.

To better understand trends for selected categories, Figure 29.6 reports the normalized frequency, calculated for three arbitrarily selected categories, as the actual number of occurrences of a particular category minus the mean of all frequencies divided by the standard deviation. The *Semantic_Web* category appeared in LAK publications in 2013 and a slight increment can be observed between 2013 and 2014. The analysis of the trend for the *Discourse_Analysis* category reveals a positive increment over the years with a remarkable increment registered in the last year. On the contrary, we observe a negative trend for the *Social_networks* category; in fact, the relevance of this category decreased substantially from 2011 to 2013, with a slightly increment in 2014.

Is There a Link Between the LD and LA Communities?

As indicated above, the analysis of authors contributing to the LAK community and the topic coverage of LAK publications provides clues about the influence of Semantic Web on researchers in the LAK community, a question of relevance to the scope of this article. Figure 29.7 shows the percentage of authors linked with either DBLP or the Semantic Web Dog Food dataset, showing a positive trend related to the increment of authors from the Semantic Web community. This can be attributed either to SW researchers publishing more strongly in the LA community or that LA researchers began publishing in SW-related venues.

To investigate this further, the links between the authors of the LAK dataset and the Semantic Web Dog Food have been exploited to determine the number of Semantic Web-related publications by LA authors. These have been measured by the *number of publications in the SWDF dataset by LA authors*. As we already know from Figure 29.2, SW conferences are not in the top 10 list of previous publications for LAK authors, but the percentage of papers published by LA authors in SW conferences shows a positive trend over the years, even if the total number reduced in 2014, as reported in Figure 29.8.

While some of these insights are hardly surprising, the ease with which they could be generated is worth highlighting: in all cases, data was fetched with a few SPARQL queries, where the previously established links between co-references across different datasets (LAK, DBpedia, DBLP, SWDF) allows the correlation of data from these different sources to answer more complex questions.

CONCLUSIONS AND LESSONS LEARNED

Applying LD principles when dealing with LA data, or any kind of data, has benefits specifically for understanding and interpreting data. As a key component of LD principles, one of the enabling building blocks is the use of global URIs for identifying entities and schema terms across the Web, which provides the foundations for cross-dataset linkage and querying, essentially creating a global knowledge graph.

In order to demonstrate the opportunities arising from adopting LD principles in LA and present some insights into the state and evolution of the LA community and discipline, we have introduced the LAK dataset, together with a set of example questions and insights. These include investigations into the composition of the LA community as well as the significant topics and trends that can be derived from the LAK dataset when considering other LD sources, such as DBLP or DBpedia, as background knowledge.

While these insights were not meant to provide a thorough investigation of the state of the LA field, they provide a glimpse into the opportunities arising from following LD principles and exploiting external data sources for interpreting data and investigating more complex research questions, which would not be feasible by looking at isolated data sources.

In this regard, a number of best practices emerge when sharing and reusing data on the Web, concerning 1) the data publishing side and 2) the data analysis side. Regarding the former, previous work (Dietze, Taibi, & d'Aquin, 2017) describes the practices and design choices applied when building and publishing the LAK dataset. Here, next to the general LD principles, we paid particular attention to designing a schema from established and well-used vocabulary terms. We considered a range of criteria, including the wide adoption of the used vocabulary terms, their coverage and match with the data model of the LAK dataset, as well as their inherent compatibility. We applied similar criteria when choosing linking candidates, such as DBLP or DBpedia, to enable more meaningful analysis of the LA community and its scientific output. While finding candidate datasets for the linking task is an inherently difficult problem, automated approaches (Ben Ellefi et al., 2016) can be applied to aid dataset providers.

While our initial analysis of the LAK dataset only provided a limited perspective on certain aspects of the LAK community and its evolution, it illustrates the ease with which particular research questions can be investigated using a well-defined and interlinked dataset, as opposed to a traditional database. More thorough studies of the LA community have been carried out as part of the LAK Data Challenge, in which researchers have been invited to develop applications aimed at providing innovative exploration of the data contained in the LAK dataset.

REFERENCES

d'Aquin, M., Adamou, A., & Dietze, S., (2013). Assessing the educational linked data landscape. *Proceedings of the 5th Annual ACM Web Science Conference* (WebSci '13), 2–4 May 2013, Paris, France (pp. 43–46). New York: ACM.

d'Aquin, M., & Jay, N. (2013). Interpreting data mining results with linked data for learning analytics: Motivation, case study and direction. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 155–164). New York: ACM.

d'Aquin, M., Dietze, S., Herder, E., Drachsler, H., & Taibi, D. (2014). Using linked data in learning analytics. *eLearning Papers*, 36. http://hdl.handle.net/1820/5814

Ben Ellefi, M., Bellahsene, Z., Dietze, S., & Todorov, K. (2016). Intension-based dataset recommendation for data linking. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, & C. Lange (Eds.), *The Semantic Web: Latest Advances and New Domains* (pp. 36–51; Lecture Notes in Computer Science, Vol. 9678). Springer.

Bizer, C., Heath, T., & Bernes-Lee, T. (2009). Linked data: The story so far. https://wtlab.um.ac.ir/images/e-library/linked_data/Linked%20Data%20-%20The%20Story%20So%20Far.pdf

Dawson, S., Gašević, D., Siemens, G., & Joksimović, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. *Proceedings of the 4th International Conference on Learning Analytics & Knowledge*(LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 231–240). New York: ACM.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H., Giordano, D., Marenzi, I., & Pereira Nunes, B. (2013). Interlinking educational resources and the web of data: A survey of challenges and approaches. *Program: Electronic Library and Information Systems*, 47(1), 60–91.

Dietze, S., Taibi, D., Yu, H. Q., & Dovrolis, N. (2015). A linked dataset of medical educational resources. *British Journal of Educational Technology*, 46(5), 1123–1129.

Dietze, S., Taibi, D., & d'Aquin, M. (2017). Facilitating scientometrics in learning analytics and educational data mining: The LAK dataset. *Semantic Web Journal*, 8(3), 395–403.

Dietze, S., Drachsler, H., & Giordano, D. (2014). A survey on linked data and the social web as facilitators for TEL recommender systems. In N. Manouselis, K. Verbert, H. Drachsler, & O. C. Santos (Eds.), *Recommender systems for technology enhanced learning: Research trends and applications* (pp. 47-77). Springer.

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. San Rafael, CA: Morgan & Claypool Publishers.

Taibi, D., Fetahu, B., & Dietze, S. (2013). Towards integration of web data into a coherent educational data graph. *Proceedings of the 22nd International Conference on World Wide Web* (WWW '13), 13–17 May 2013, Rio de Janeiro, Brazil (pp. 419–424). New York: ACM. doi:10.1145/2487788.2487956