Towards Analysing the Scope and Coverage of Educational Linked Data on the Web

Davide Taibi, Giovanni Fulantelli
National Research Council of Italy
Institute for Educational Technologies
Via Ugo La Malfa 153 - 90146 Palermo, Italy
{davide.taibi, giovanni.fulantelli}@itd.cnr.it

Stefan Dietze, Besnik Fetahu
L3S Research Center
Appelstraße 9A
30176 Hannover, Germany
{dietze, fetahu}@l3s.de

ABSTRACT
The diversity of datasets published according to Linked Data (LD) principles has increased in the last few years and also led to the emergence of a wide range of data suitable in educational settings. However, sufficient insights into the state, coverage and scope of available educational Linked Data seem to be missing, for instance, about represented resource types or domains and topics. In this work, we analyse the scope and coverage of educational linked data on the Web, identifying the most popular resource types and topics, apparent gaps and underlining the strong correlation of resource types and topics. Our results indicate a prevalent bias to-wards data in areas such as the life sciences as well as computing-related topics.

Categories and Subject Descriptors
I.2.4 [**Knowledge Representation Formalisms and Methods**]; H.3.1 [**Content Analysis and Indexing**];

General Terms
Design, Measurement, Experimentation

Keywords
Dataset profile, Linked Data for Education, Linked Data Explorer.

1. INTRODUCTION
The diversity of datasets published according to Linked Data (LD) principles [4] has increased in the last few years and also led to the emergence of a wide range of data suitable in educational settings. These include open educational resource metadata, statistical data about the educational sector, video lecture metadata or university data about courses, research or experts [6]. Initial efforts to collect and catalog such datasets have been made through initiatives such as the LinkedUp Data Catalog[1] or related community initiatives[2].

However, sufficient insights into the state, coverage and scope of available educational Linked Data seem to be missing. Here, particular questions about the represented resource types (such as, resource metadata or information about organisations or people) and topics, are of crucial relevance to shape a better understanding about the state of educationally relevant Linked Data on the Web [5], [10]. Also identifying a dataset containing resources related to a specific topic is, at present, a challenging activity. Moreover, the lack of up-to-date and precise descriptive information has exacerbated this challenge. The mere keyword-based classification derived from the description provided by the dataset owner is not sufficient, and for this reason, it is necessary to find new methods that exploit the characteristics of the resources within the datasets to provide useful hints about topics covered by datasets and their subsequent classification.

In this direction, authors in [1] proposed an approach to create structured metadata to describe a dataset by means of topics, defined as DBpedia categories, where a weighted graph of topics constitutes a dataset profile. Profiles are created by means of a processing pipeline that combines techniques for dataset resource sampling, topic extraction and topic ranking. Topics were extracted by using named entity recognition (NER) techniques, where topics are ranked, respectively weighted, according to their relevance using graph-based algorithms such as PageRank, K-Step Markov, and HITS.

The limitations of such an approach are related mainly to the following aspects. First, the meaning of individual topics assigned to a dataset can be highly dependent on the type of resources they are attached to. Also, the entire topic profile of a dataset is hard to interpret if categories from different types are considered at the same time. As an example of the first issue, the same category (e.g. "Technology") might be associated to resources of very different types such as "video" (e.g. in the Yovisto dataset[3]), “research institution"(e.g. in the CNR dataset[4]), or medical learning resources (e.g. the dataset of the mEducator project [9]). Concerning the second issue, the single topic profile attached for instance to bibliographic datasets (such as: the LAK dataset [7] or Semantic Web Dog Food[5]) - in which people (“authors”), organisations ("affiliations") and documents (“papers”) are represented – is characterised by the diversity of its categories (e.g. DBpedia categories: Scientific_disciplines, Data_management Information_science but also Universities_by_country, Universities_and_colleges). Indeed, classification of datasets in the LD Cloud is highly specific to the resource types one is looking at. While one might be interested in the classification of "persons"

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

[2] https://www.w3.org/community/opened

[3] http://www.yovisto.com/

[4] http://data.cnr.it/

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

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741741.

listed in one dataset (for instance, to learn more about the home countries of authors in DBLP[6]), another one might be interested in the classification of topics covered by the documents (for instance disciplines of scientific publications) in the very same dataset.

In this paper, we aim at providing a systematic assessment of educational Linked Data which considers both, represented topics as well as resource types, and their correlations. The approach we propose overcomes the limitations described above by considering the topic profiles defined in [1] in the context of the resource types they are associated with. However, the schemas adopted by the datasets of the LD cloud are heterogeneous, thus making it difficult to compare the topic profiles across datasets. While there are many overlapping type definitions representing the same or similar real world entities, such as "documents", "people", "organization", type-specific profiling relies on type mappings to improve the comparability and interpretation of types and consequently, profiles. To this aim, the explicit mappings and relations declared within specific schemas (for instance, *foaf:Person* being a subclass of *foaf:Agent*) as well as across schemas (for instance through *owl:equivalentClass* or *rdfs:subClassOf* properties) are crucial. While relying on explicit type mappings, we have based our work on a set of datasets where explicit schema mappings are available from earlier work [2]. This includes education-related datasets identified by the LinkedUp Catalog in combination with the dataset profiles generated by the Linked Data Observatory[7]. While the latter provides topic profiles for the majority of LD datasets, the LinkedUp Catalog contains explicit schema mappings which were manually created for the most frequent types in the LinkedUp Catalog. Using these resources, we provide a broad overview of the coverage, scope and gaps of available Linked Data to be used in educational settings.

The next Section provides an overview of the methodology applied to shape the landscape of the educational linked data. The results of our analysis are discussed in Section 3. In particular, a network analysis tool has been used to provide a resource type-specific overview of the categories as well as the resource types associated with the datasets in the LinkedUp Catalog.

2. METHODOLOGY

In the framework of the study presented in this paper, the research questions of interest can be summarised as follows:

Q1: Which types and topics are covered by existing educational Linked Data?

Q2: What are the central topics covered for particular types, (e.g. Open Educational Resources metadata)?

Q3: Are certain topics underrepresented for certain types, or vice versa?

These research questions focus on three key elements: *datasets*, *topics* and *resource types*.

Since resource types can only be considered if resource type mappings are available, we exploited such mappings from the LinkedUp Catalog metadata dataset[8].

Topic profiles are taken from the dataset of topic profiles[9], further described in [3], where topic annotations in the form of DBpedia categories are provided for the majority of LD datasets. A topic profile connects a dataset with the topics extracted from the analysis of resource samples. Since topics are ranked, a topic profile can be seen as a weighted dataset-topic graph. As such, a topic profile provides a comprehensive overview of the topic coverage of individual datasets. Analysed across a specific set of datasets - as carried out in this work - topic profiles provide insights into the coverage of such a set of datasets.

While topic annotations are obtained from analysing resources of a particular type, the semantics of the topic can best be interpreted when considering the type of the resource. As an example, if the topic "Biology" is associated to a *foaf:Document* resource it is likely referred to a scientific paper related to biological aspects.

Table 1: List of selected datasets

Dataset	Total data	
	#Types	#resources
asn-us	29	7494200
colinda	21	17006
data-cnr-it	120	485977
data-open-ac-uk	134	386291
education-data-gov-uk	99	315632
educationalprograms_sisvu	27	104238
gesis-thesoz	9	48532
hud-library-usagedata	6	904747
l3s-dblp	6	15514
lak-dataset	14	13688
linked-open-aalto-data-service	22	373553
morelab	13	244
open-courseware-consortium-metadata-in-rdf	4	22850
organic-edunet	1	11093
oxpoints	142	73655
publications-of-charles-university-in-prague	258	14324
seek-at-wd-ict-tools-for-education-web-share	556	13502
unistat-kis-in-rdf-key-information-set-uk-universities	35	371737
universitat-pompeu-fabra-linked-data	39	5778
university-of-bristol	15	240179
yovisto	8	549986

In the case the "Biology" topic is associated to a *foaf:Organization* resource, it is likely referred to a Biology department of a university.

Since our work considers the investigation of both, topics and types, our research was limited to 21 datasets, which were the ones existing in both collections, i.e. where both topic profile and resource type mapping annotations were available. The complete list of selected datasets is shown in Table 1.

[6] http://dblp.l3s.de/

[7] http://data-observatory.org/lod-profiles/

[8] http://datahub.io/dataset/linkedup-catalogue-of-educational-datasets

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

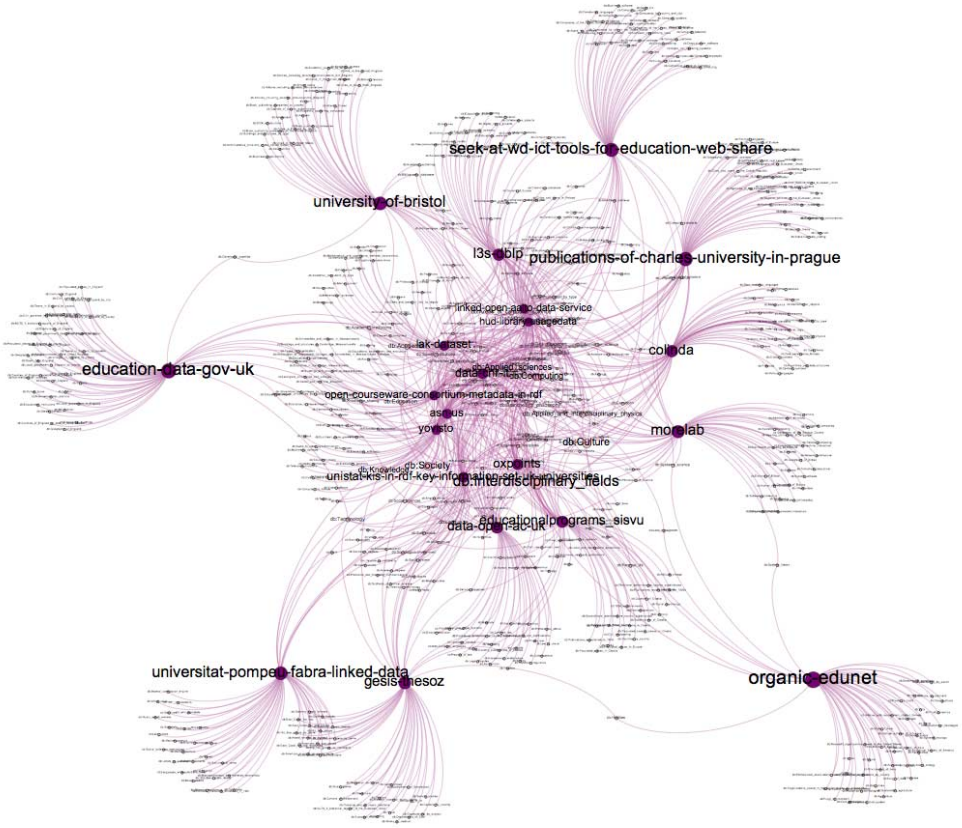

Figure 1: Dataset and categories network

The analysis of the relationships between datasets, topics and resource types - aimed at providing a response to the research questions posed at the beginning of this section - was undertaken by exploiting network analysis theories and methods. Graph centrality measures have been used to identify the relevance of the nodes. In particular, the *betweenness centrality* measure has been used. Despite other measures that help measuring the importance of the nodes based on their topological position, the betweenness centrality of a node is calculated by considering the number of the shortest paths from all pairs of nodes that pass through the node. Indeed, the connections between the three investigated notions can be represented by networks, in which the elements are nodes and their relationship are edges. Specifically the analysis of the relationships has been conducted by considering:

- the network representing the relationships between datasets mediated by categories
- the network representing the relationships between datasets mediated by resource types
- the network representing the relationships between resource types mediated by categories

These networks have been represented by using the Open Source software Gephi[10]. Exploiting the insights gained from such networks, we can identify the particular type/topic coverage of educational LD datasets, corresponding gaps, and the correlation of educational resource types and topics.

3. ANALYSING THE EDUCATIONAL LINKED DATA LANDSCAPE

In this section, we describe the results of our dataset landscape analysis.

3.1 Analysing topic coverage - the dataset-category-graph

Representing datasets and categories, i.e. topics, as a weighted graph allows us to analyse the topic coverage of assessed datasets and their proximity topic-wise. In particular, a dataset is connected with the corresponding category depending on its topic profile. Indirect relationships among datasets emerge through shared or connected categories.

Table 2 reports the list of the top ten most connected categories in the datasets under investigation by taking into consideration the

[10] http://gephi.github.io/

number of resources. As stated in section 2, in the dataset profile each topic is a DBpedia category, though we omitted the DBpedia namespace (http://dbpedia.org/category/) from the listing. The number of datasets sharing the specific category is also reported.

The categories reported in Table 2 highlight the heterogeneity of the dataset resources: categories representing actual disciplines (such as *Biology*, *Computing*, as extracted from Open Educational Resources or video lectures) as well as categories related to institutions (such as *Academic_institutions*, *Academic_disciplines*) are represented in the list. This overview already demonstrates the strong impact of the resource type (eg *foaf:Document* or *foaf:Organisation*) on the associated categories, an observation which motivated parts of the following investigations and an explorative browser described in [8]. The network of dataset and categories is shown in Figure 1[11].

Table 2. Most frequent categories in educational Linked Data

Category	# Datasets	# resources
Applied_sciences	19	3581
Computing	16	2778
Academic_disciplines	19	2328
Biology	16	2068
Digital_technology	12	2012
Education	14	1855
Academia	15	1668
Academic_institutions	14	1625
Interdisciplinary_fields	16	1574
Applied_disciplines	18	1368

3.2 Resource Type coverage - the dataset-type-graph

To provide an overview of represented resource types, we build on previous work in [2] and generate a dataset-type-graph, where nodes are resource types and datasets, and an edge connects a dataset with a resource type if the dataset contains resources of that type (Figure 2). In the network of Figure 1 the resource type is not considered, thus two datasets can be connected even if they are collecting different typologies of resources such as information about institutions, learning materials or scientific publications. In principle, the type of the resources plays a key role to guide the exploration of the datasets and furthermore it heavily influences the connections between datasets. For this reason, in this study we introduced a new layer of analysis by considering the type of the resources within the datasets. Therefore, the influence of the resource types in the relationships between datasets has been investigated.

In order to improve the analysis of the relationships between datasets and resource types, both *explicit* and *implicit* mappings have been introduced. As *explicit* mapping, we consider the relationships that can be inferred and are explicitly declared in the vocabulary used in the datasets. In addition, in the context of the LinkedUp project[12] a set of additional mappings has been introduced which link equivalent or overlapping types through standard OWL and RDF predicates, such as, *owl:equivalentClass* or *rdfs:subClassOf*. A detailed description of the process that has led to the definition of these mappings is described in [2]. Table 3 reports the ten resource types most shared by the 21 datasets of the under investigation.

Table 3. Most frequent resource types according to their representation in the datasets

Resource Type	# Datasets
foaf:Agent	**14**
foaf:Person	5
foaf:Organization	3
aiiso:Institution	3
foaf:Agent	3
aiiso:Department	2
foaf:Document	**12**
foaf:Document	5
bibo:Article	2
bibo:Book	2
bibo:Document	2
swrc:Document	2
swrc:InProceedings	2
aiiso:KnowledgeGrouping	**7**
aiiso:Course	3
aiiso:Module	2
courseware:Course	2
skos:Concept	**6**
skos:Concept	4
geo:SpatialThing	**4**
c4dm:Event	**3**
void:Dataset	**3**

In this table, the resource types have been grouped by considering the relationships defined by implicit and explicit mappings. For this reason, for instance, *foaf:Person* and *foaf:Organization* types are not represented since they are subclasses of *foaf:Agent*. the most represented resource types are related to *foaf:Document* (since there are datasets collecting scientific and academic publications), *foaf:Agent* (some of the datasets under investigation contain information about organizations, institutions and people) and *aiiso:KnowledgeGrouping*, since this class represents resources related to courses, learning modules, and so on.

[11] Hi-res version of this image is available at: http://data-observatory.org/led-explorer/lile_fig_1.svg

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

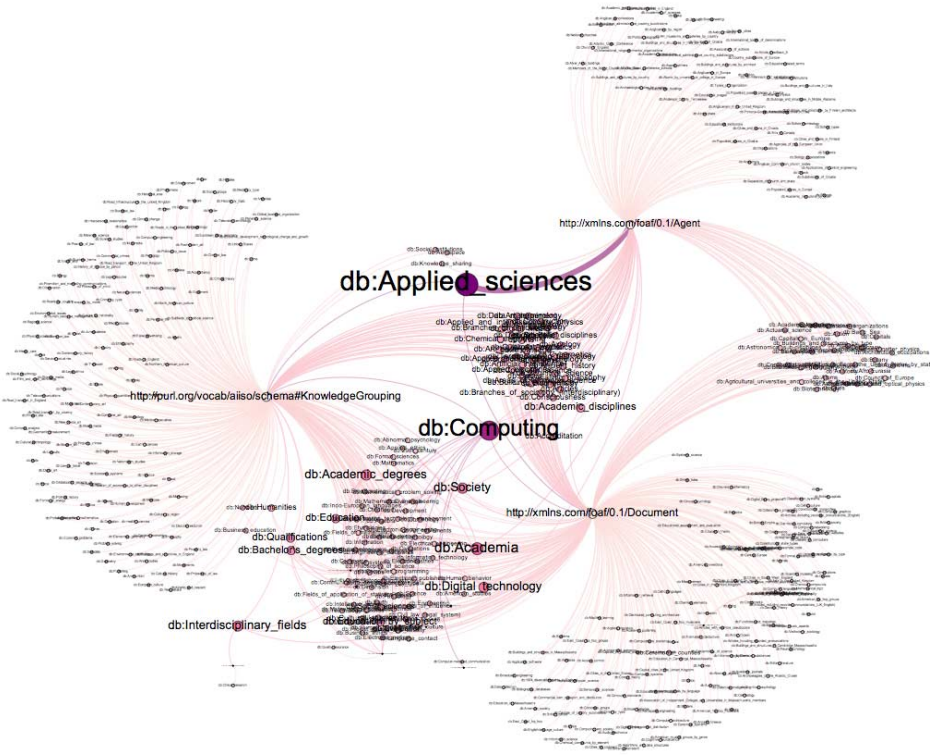

Figure 2. The network of resource types and categories

Type mappings across all involved datasets link "documents" of all sorts to the common *foaf:Document* class, "persons" and "organisations" to the common *foaf:Agent* class, and course and module to the *aiiso:KnowledgeGrouping* class. In table 3 the resource types that appear in only one dataset are not reported.

3.3 Type-Topic Correlation

As shown in Figure 2[13], the resource type has a strong impact on the nature and semantics of the associated categories. While actual knowledge resources, such as OER, tend to be linked to explicit domains or disciplines, such as *Biology* or *Computer Science*, the range of categories for persons and organisations is of entirely different nature.

While topics/categories are always linked to particular resources and their types, the joint analysis of both types and topics is of crucial importance to enable a better understanding of educational Linked Data. Considering the resource types associated with each topic in the dataset topic profile graph, it has been possible to create a network in which the resource types have been connected with the categories they are related to.

Table 4 reports the most representative categories related with the most connected resource types in the LinkedUp catalog. In order to enable a better distinction, we particularly consider the highest level types, e.g. *foaf:Agent* rather then *foaf:Person* and *foaf:Organisation*.

Table 4. Most frequent categories for most frequent resource types

foaf:Document		**foaf:Agent**		**aiiso:KnowledgeGrouping**	
Applied_sciences	1164	Applied_sciences	1522	Digital_technology	1393
Biology	680	Academic_institutions	533	Computing	1262
Academic_disciplines	656	Academic_disciplines	823	Society	1011
Branches_of_philosophy	624	Educational_organizations	533	Interdisciplinary_fields	793
Chemistry	604	Types_of_organization	523	Education_by_subject	789
Areas_of_computer_ science	593	School_types	520	Academia	717
Education	591	Schools	520	Academic_disciplines	688
Artificial_intelligence	581	Organizations	520	Education	653
Computing	548	Educational_institutions	520	Applied_sciences	648
Branches_of_psychology	548	Educational_buildings	516	Qualifications	591

[13] Hi-res version of this image is available at: http://data-observatory.org/led-explorer/lile_fig_2.svg

Table 4 provides evidence that the resources related to persons and organizations (*foaf:Agent*) are more connected to physical places and locations, while resource types related to actual documents (*foaf:Document*) or courses (*aiiso:KnowledgeGrouping*) are more related to learning topics. For the latter, we observe a strong bias towards topics relating to Computer Science and the Life Sciences. This observation correlates with the general intuition that such topics are also stronger represented in the Linked Open Data cloud in general and might lead to additional research into how to resolve such a topic bias in the future.

4. CONCLUSIONS

As demonstrated in this paper, topic profiling of datasets needs to take into account the association between resource types and topics. Only the joint consideration of types and topics allows the non-ambiguous interpretation of topic annotations of datasets. Our analysis uncovers an inherent topic bias of educational resources represented in datasets, usually focused on disciplines related to *Computer Science* and *Life Sciences*, where for instance, social sciences appear to be underrepresented. The analysis of the resource types highlights that documents, such as scientific publications and books, are more represented than videos, while other media which are also used in educational contexts, such as images, are scarcely represented.

In work [8], strongly related to this, we presented an explorative interface which allows to browse and explore educational Linked Data by considering type and topic annotations. While our current work so far did not study the relationships emerging from the inherent relatedness of DBpedia categories as captured by the DBpedia category graph, future work will explore these relationships. For instance, if the dataset D_1 refers to category C_x and dataset D_2 refers to category C_y, the path between C_x and C_y in the DBpedia category graph (e.g. C_x might be a subcategory of C_y) might also hint at additional connections. In this sense, an additional set of relationships will be introduced, allowing for more sophisticated dataset exploration. Finally, future work will also aim at establishing to what extent the similarity of the topic distribution of datasets can serve as an indicator of the similarity of their respective, disparate resource types.

5. ACKNOWLEDGMENTS

This work has been partially supported by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No 317620 – LinkedUp project (http://linkedup-project.eu/).

6. REFERENCES

[1] Fetahu, B., Dietze, S., Nunes, B. P., Taibi, D., Casanova, M. A.. 2013. Generating structured Profiles of Linked Data Graphs. *In Proceedings of the 12th International Semantic Web Conference (ISWC2013)*,(Sydney, Australia, 2013).

[2] D'Aquin, M., Adamou, A., Dietze, S. 2013. Assessing the Educational Linked Data Landscape. *In Proceedings of ACM Web Science 2013 (WebSci2013)*, Paris, France, May 2013.

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

[4] Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data - the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1-22. doi:10.4018/jswis.2009081901

[5] Dietze S., Yu H. Q., Giordano D., Kaldoudi E., Dovrolis N., Taibi D. 2012. Linked Education: interlinking educational Resources and the Web of Data. *ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications*.

[6] Dietze S., Sanchez-Alonso S., Ebner H., Yu H. Q., Giordano D., Marenzi I., Pereira Nunes B. (2013). Interlinking educational resources and the web of data: a survey of challenges and approaches. *Emerald Program: electronic library and information systems*, 47(1), 60-91. doi: 10.1108/00330331211296312.

[7] Taibi, D. and Dietze, S. 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).

[8] Taibi, D., Dietze, S., Fetahu, B., Fulantelli, G., Exploring type-specific topic profiles of datasets: a demo for educational linked data, in Poster & System Demonstration Proceedings of 13th International Semantic Web Conference (ISWC2014), Riva Del Garda, Italy, October 2014.

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

[10] Taibi, D., Fulantelli, G., Dietze, S., Fetahu, B. 2013. Evaluating Relevance of Educational Resources of Social and Semantic Web. In *Scaling up Learning for Sustained Impact*, 637-638. Springer Berlin Heidelberg.