A Linked Dataset of Medical Educational Resources

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

Hong Qing Yu[a*], Stefan Dietze[b], Davide Taibi[c], Daniela Giordano[d], Eleni Kaldoudi[e] and John Domingue[a]

[a]*Knowledge Media Institute, The Open University, United Kingdom*
[b]*L3S Research Centre, Leibniz University, Hanover, Germany*
[c]*Italian National Research Council, Institute for Educational Technologies, Italy*
[d]*University of Catania, Dipartimento di Ingegneria Elettrica, Elettronica e Informatica, Catania, Italy*
[e]*Democritus University of Thrace, Greece*

Abstract. With sharing and reusing, educational resources become increasingly important for enhancing learning and teaching experiences, particularly in medical educational domain since these resources are expensive to re-produce. In respect to this, many efforts have been applied to federate the resources to achieve the sharing and reusing goals, which led to a fragmented landscape of competing metadata schemas, such as IEEE LOM or OAI-DC, and interface mechanisms, such as OAI-PMH or SQI. However, the major issue of educational resource federating is the heterogeneity challenge of metadata and data. In this paper, we illustrate a medical educational dataset (mEducator Linked Educational Resources dataset) that is published as part of the Linked Open Data cloud following Linked Data principles. The dataset contains educational resource metadata federated from ten different (medical) educational institutes together with enriched links to related information by using Linked Data techniques and datasets. We introduce a Semantic Web Service based data extracting mechanism that is exploited for services and data integration to address heterogeneous metadata problems. The paper also discusses the dataset accessing APIs, statistics and existing applications of using the mEducator dataset.

Keywords: Linked Data, medical educational resource, Semantic Web, Web services, Web APIs, learning repository

1 Introduction

Recently, we started to witness increasing usage of Linked Data (LD) [1] technologies for publishing educational resources metadata in order to internally and externally link related educational resources.

However, the federation of existing educational resources from different institutions and domains is still a big challenge.

Although there is already a large amount of educational data available on the Web via proprietary and/or competing schemas and interface mechanisms, the main challenges are to (a) start adopting LD principles and vocabularies while (b) leveraging on existing educational data available on the Web via non-LD compliant means and (c) multiple levels of heterogeneous issues such as educational resource metadata schema and data retrieving interfaces.

In this paper, we illustrate a linked medical educational resource dataset (the mEducator dataset) that has been published in the Linked Open Data (LOD) cloud as one important outcome of the mEducator project[1]. By applying a Linked Service [6] based data federation framework, the mEducator dataset has integrated medical educational resources from 10 heterogeneous educational repositories. The mEducator dataset is daily updated by dynamically harvesting the educational resource metadata through the repository Web APIs/Services. The framework also provides the capability of dealing with heterogeneous metadata mediation, data enrichment and clustering, and flexible data extension.

The remainder of the paper is organised as follows: section 2 explains the mEducator OWL-based data schema and the accessing endpoint; section 3 introduces the overall Semantic Web Service based architecture of the Linked Service framework to illustrate the data integration methodology and implementation; section 4 explains the data enrichment and clustering techniques; the mEducator Linked Data has been used and integrated into five medical educational related Web and mobile applications currently. The application examples and

[1] http://www.meducator.net

evaluation statistics will be discussed in section 5; finally, section 6 draws a conclusion and discusses the future work.

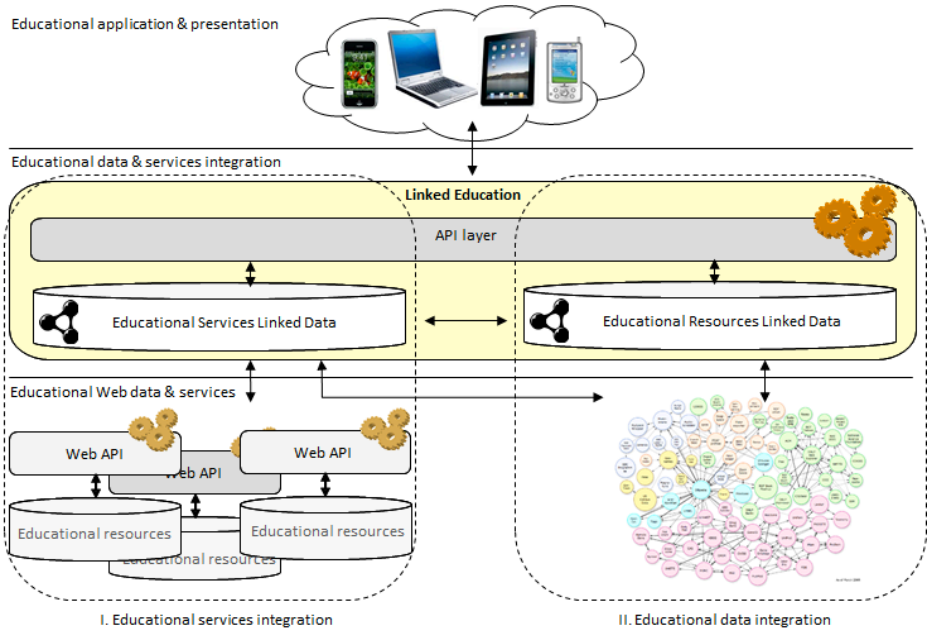

Figure 1 Overview of the data generation and maintaining process

2 Data modelling schema and data availability

2.1. The mEducator schema

In order to define the suitable medical learning resource metadata modelling schema, three widely used metadata standards (e.g. IEEE LOM [1] and Dublin Core [3]) were studied, in which learning object modelling concepts, properties, vocabularies and their relations are investigated. Based on the studies, the mEducator schema is defined by integrating those commonly used concepts, vocabularies and properties under one unified model and extending the vocabularies with more properties that facilitate the medical educational resource sharing and interactive capabilities. The mEducator metadata schema[2] [5] consists of title, identifier, the language in which it is created, the language of its metadata, the date of resource creation, the date of metadata creation the resource author, the author of the metadata, a brief description, a technical description, its IPR license, citation and any further information that certifies its quality. In addition, the mEducator metadata schema is proposed to provide pedagogical values such as educational objectives, assessment methods, teaching instructions, educational level, educational prerequisites and educational outcomes. These additional metadata vocabularies are also used to provide keyword-based resource descriptions and to describe the using discipline and discipline specialty in the medical domain to which a resource relates.

2.2. Data storage and accessing

All extracted and generated educational metadata are eventually stored in a dedicated RDF store as the mEducator dataset[3]. This store is implemented using Sesame/BIGOWLIM[4] and is compliant with the mEducator schema. Each educational resource metadata entity described using the mEducator resource schema owns a unique and dereferenceable URI, such as http://purl.org/meducator/resources/25a8c581-66d7-4186-9411-f9f0f783463e. In addition, a set of dedicated REST APIs is implemented to enable client applications to query, store and retrieve the metadata in the RDF store (an authentication key is required to use the APIs).

To foster different application development requires, these REST APIs contain different types of queries:

- SPARQL [18] endpoint query[5]
- keyword query[6]
- property-based keyword query[7]

[2] http://www.purl.org/meducator/ns/

[3] http://ckan.net/package/meducator

[4] http://www.ontotext.com/owlim/

[5] http://meducator.open.ac.uk/resourcesrestapi/rest/meductor/auth/sparql?query=

[6] http://meducator.open.ac.uk/resourcesrestapi/rest/meducator/auth/keywordsearch?keyword=

- rdfs:seeAlso-based keyword query[8]
- identifier-based property query[9]
- push method for inserting an education resource into the RDF repository[10]

3 Data sources, generation and maintenance

3.1. Overview of the process and data sources

The proposed architecture, which has been introduced in [19] and fully described in [20], includes three layers (see Figure 1): Educational (Web) data and service layer, Educational data and service integration layer and Educational application and presentation layer. The data sources are originally derived from 10 medical educational metadata repositories (see Table 1) and legally conducted by the mEducator project and the Open Source licenses.

Service	Base URI
Pubmed	http://www.pubmedcentral.nih.gov/oai/oai.cgi
University of Catania OER	http://151.97.9.184/WebTraceRDFApi/Default.aspx?
Biblioteca Digital de Teses e Dissertações da UFRN	http://bdtd.bczm.ufrn.br/tedesimplificado/tde_oai/oai2.php
Aristotle University of Thessaloniki, Moodle repository	http://kedip.med.auth.gr/meducatormoodleapi/index.php
Open Research Online (ORO), Open University, UK	http://libeprints.open.ac.uk/cgi/oai2
Medical learning resources, Technical University of Cluj-Napoca	http://dataserver.mediogrid.utcluj.ro/adnotare/cluj_endpoint.php
Educational resources of EUREKA project (Canada)	http://eureka.ntic.org/oai-pmh.php
Learning resources, University of Helsinki (Finland)	https://helda.helsinki.fi/dspace-oai/request
Linked Data Learning resources, University of Helsinki (Finland)	https://helda.helsinki.fi/mEducatorWebapp/RDFEndpoint
Nice University educational resources (France)	http://revel.unice.fr/oai/oai2.php

Table 1 The list of Data source repositories

[7]http://meducator.open.ac.uk/resourcesrestapi/rest/meductor/auth/propertysearch?property=&value=

[8]http://meducator.open.ac.uk/resourcesrestapi/rest/meductor/auth/eidsearch?id=

[9]http://meducator.open.ac.uk/resourcesrestapi/rest/meductor/auth/search?ids=&properties=

[10]http://meducator.open.ac.uk/resourcesrestapi/auth/rest/meducator/

3.2. Data extracting via Semantic Web Services

As shown in Figure 1, the educational repositories are integrated through Web APIs, and data are harvested through dynamically invoking and lifting the responses of the Web APIs.

In order to annotate and integrate learning resource repository accessing APIs, we exploit two well-integrated technologies: iServe[11] [6] and SmartLink[12] [8] (Both iServe and SmartLink adopt LD principles to expose services and Web APIs). iServe is a framework that provides service modelling vocabularies and related tools for dynamic discovering and invoking services and Web APIs. SmartLink is a LD-based Web application[13] that handles two different kinds of service annotations separately, namely functional (iServe Minimum Service Modelling vocabularies) and non-functional service annotations stored in dedicated RDF stores. In addition, SmartLink allows browsing existing descriptions within the RDF service metadata repositories. Our services integration approach consists of the following steps:

1. annotating services and publishing these annotations through the SmartLink Web interface;
2. iServe discovers suitable services for keeping the dataset updated at runtime using REST APIs;
3. invoking services and lifting heterogeneous service responses into the mEducator schema through the iServe invocation engine [15].

Figure 2 shows an example of lifting XML data to RDF data. Lifted RDF data will be cached and integrated into the mEducator repository.

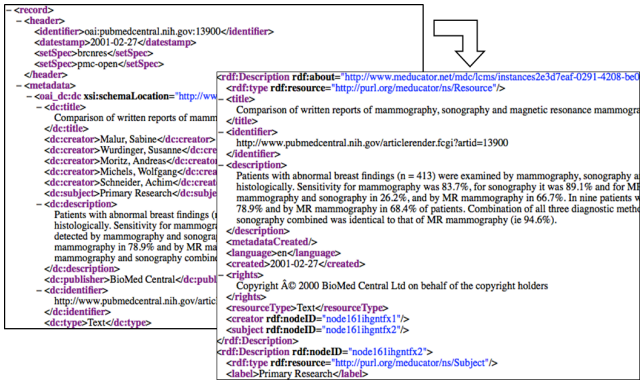

Figure 2 Lifting XML metadata description to the mEducator RDF description

3.3. Manual data annotations inserted via REST API

The push method mentioned in Section 2.2 deals with the data inserting requests from client applications that allow application users to describe medical learning resources in RDF style directly. Meanwhile, the

[11] http://iserve.kmi.open.ac.uk/

[12] http://smartlink.open.ac.uk

[13] http://smartlink.open.ac.uk/smartlink

integrated heterogeneous educational data that retrieved from previous data extracting process can be re-annotated as an RDF resource and stored in the mEdcuator dataset repository.

4 Data enrichment and clustering

4.1. Data enrichment for generating external links

Data enrichment is implemented in two ways (a) as automated mechanism whenever new data are pushed to the RDF store and (b) also as semi-automated approach where users are provided with suggestions of related terms from which they can select suitable ones as part of a particular end user application. While the first approach makes use of DBpedia exclusively, resulting in large numbers of automatically retrieved references to DBpedia resources, the second approach makes exclusive use of the BioPortal API exclusively.

Linking resources to shared LOD vocabularies serves two main purposes. Firstly, it enhances the metadata of individual resources with additional knowledge. Secondly, it provides a means to identify correlations between individual resources that share the same external references.

External Source	Number of distinct Terms	Percentage
DBPedia	509	93.39
Medical Subject Headings	11	2.02
SNOMED Clinical Terms	11	2.02
Health Level Seven	4	0.73
Galen	2	0.37
MedDRA	2	0.37
LOINC	1	0.18
MedlinePlus Health Topics	1	0.18

Table 2 Distribution of the sources of the used terms.

The number of enrichment triples in the data store is 1352. **Table 2** shows that the enrichment involves a total of 509 distinct terms from DBpedia. The average number of enrichments per enriched resource is 4.5 (min=1, max=42). Apparently, there is a large number of enrichments obtained via the automated enrichment based on the DBpedia Spotlight API, while the semi-automated approach via the BioPortal API provides a higher diversity – data from different vocabularies such as MESH and SNOMED are used – but only a very limited amount of overall enrichments because it requires manual intervention and pre-selection of suggested terms.

4.2. Data clustering for adding internal links

The clustering functionalities have been integrated in the RDF store to allow the interlinking of resources originating from different repositories. In particular, the whole clustering process has three main steps: content indexing, creation of similarity matrix and clustering. The first step parses the RDF fields of the selected subsets of the unifying metadata schema, including those that contain descriptive free text, and creates a matrix, namely, Doc-Term (DT). The DT matrix contains the frequency of each term in a given resource and consolidates information about term co-occurrences within the same resource and across all resources, which is done according to a method that weighs less common terms and takes the neighboring terms into account also as term context [7]. The second step is based on the DT matrix to create the similarity matrix S by assigning a similarity score among the resources. The S matrix then is used to classify the resources, by applying a clustering algorithm. Our implementation supports both clustering based on Kohonen maps and Aggregative clustering.

To treat the clustering results consistently with the LD approach, and integrate this information in the unifying mEducator metadata schema, a separate RDF schema has been defined (see **Figure 3** for an excerpt of the relative instance). In this schema, the classes and properties are defined to fully describe not only the clustering results, but also the complete rundown of the clustering process itself. It adopts the Provenance Vocabulary[14] that contains all the necessary properties and classes to describe the provenience of data from the Web. In the clustering schema, the data includes, but is not limited to, the clustering processing initiator agent (human, or software), the features used for indexing, the algorithm used and its parameters, the linked resources and clusters.

Figure 3 An RDF instance of a cluster (large bounding box) and the relative reference in the mEducator resource's instance (small bounding box).

The clustering process receives input mEducator resources from all the subscribed repositories in the RDF/XML format. The process initiator selects the desired fields that should be taken in consideration in the clustering. The output of the system is another RDF/XML file that is uploaded to all queried repositories.

5 Data Usages

Currently, five educational Web applications[15] have been developed interacting with the mEducator Linked dataset. For example, the data and services integration APIs and datasets presented in the previous sections are

[14] http://trdf.sourceforge.net/provenance/ns.html
[15] http://www.meducator3.net/

fully integrated in the Metamorphosis+[16] environment, which merges the paradigms of semantic and social web for sharing linked educational resources. MetaMorphosis+ realises the educational application and presentation layer. It allows viewing, management and annotation of the educational resource metadata that are retrieved via the APIs provided by the educational services and data integration layer. Meanwhile, an Android mobile application is developed for learners to search and browser medical educational resources through Smartphone. These applications have started to serve medical students and research staff for evaluations.

6 Statistic Analysis

By the date of 1st May 2012, the mEducator Linked Educational Resources RDF repository contains in total 23876 triples, of which 10206 directly refer to a total 375 distinct educational resources. The average number of triples per educational resource is 27, ranging from a minimum of 6 to a maximum of 68. **Figure 4** provides an overview of the heterogeneity of the origin of individual resource metadata.

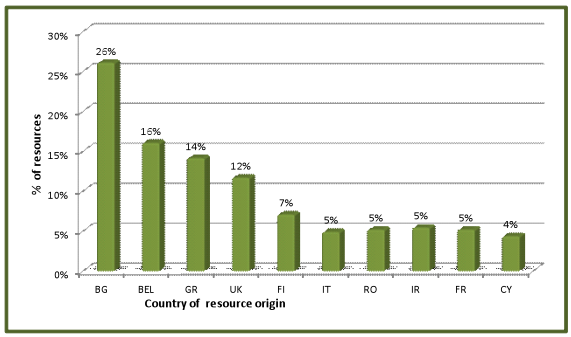

Figure 4 Number of resources (%) per country of origin (based on country of origin of contributing institution).

In addition, **Figure 5** shows that even though the metadata imported from external stores usually is very limited, often covering only less than three properties (e.g. title, description and resource location), based on our automated and semi-automated enrichment techniques, substantially large numbers of properties are provided for the majority of resources, where all resources have a minimum of 5 described properties.

Table 3 provides an overview of property usage frequency across educational resources. Properties are listed in the order of decreasing frequency. A more detailed description of individual properties is given in [17]. Data collected in the RDF store makes use of a set of external LOD vocabularies/datasets. In detail, 545 distinct terms have been associated based on the automatic and semi-automatic methodologies described in Section 4. **Table 3** shows the distribution of the terms used by external sources, and the percentage of usage of each source.

[16] http://metamorphosis.med.duth.gr/

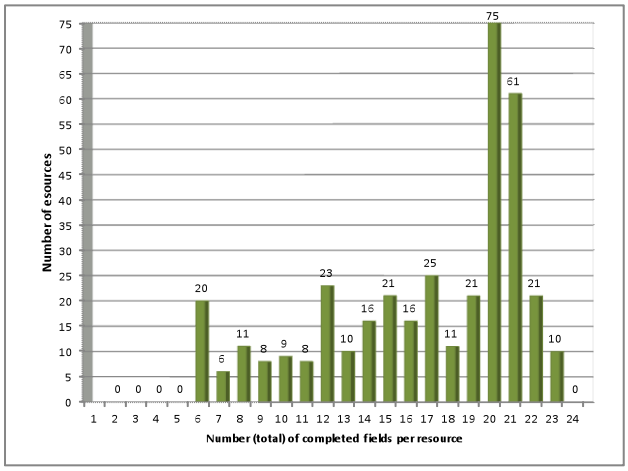

Figure 5 Frequency of total number of completed fields per resource (excluding fields that pertain only to repurposed resources).

Property	%	Property	%
mdc:creator	100	mdc:hasEnrichmentContext	79.2
mdc:description	100	mdc:rights	77.9
mdc:metadataCreated	100	mdc:discipline	76.5
mdc:metadataCreator	100	mdc:technicalDescription	68.3
mdc:title	100	mdc:educationalContext	66.7
mdc:created	95.7	mdc:educationalLevel	64
mdc:metadataLanguage	95.5	mdc:educationalObjectives	58.9
mdc:citation	94.4	mdc:teachingLearningInstructions	57.9
mdc:language	89	mdc:educationalOutcomes	51.2
mdc:resourceType	88.6	mdc:educationalPrerequisites	49.3
		mdc:disciplineSpeciality	45.6
		mdc:assessmentMethods	40.8

Table 3 mEducator schema property usage frequency across educational resources in the dataset

Table 4 lists the properties used to trigger the enrichment of the resources, and the number of enrichments obtained. It is noted that only the three free text properties are currently considered during the enrichment process.

Enriched Property	Number of enrichments	Percentage
mdc:description	733	54.2
mdc:title	327	24.2
mdc: educationalOutcomes	292	21.6

Table 4 Distribution of external DBpedia terms to resource properties

7 Conclusion and future work

In this paper, we introduced the mEducator medical learning resource dataset that is fundamentally built on Semantic Web Service technologies. The dataset is published following Linked Data principles to support Web-scale interoperability between educational resources,

that is, educational services as well as data. Linked Data are adopted to describe both services and data, which allow the integration of existing educational repositories at both the service and the data levels. By exposing educational resources via Linked Data principles, we leverage on the wealth of existing datasets and vocabularies, so that internal links between educational data and resources are generated. We have introduced a set of implemented integration approaches, resulting RDF datasets, APIs, and applications (e.g. MetaMorphosis+ and Android app) that provide open environments for medical education. One of the longer-term goals is to use the principles described in this paper to establish a unified entry point to well-interlinked educational datasets on the Web.

Future extension work will focus on two major areas of (1) investigating methods to enable integrate data from other educational domains; (2) extending the framework with additional open repositories and data stores to further showcase and evaluate our services and data integration approach.

Acknowledgments. This work is partly funded by the mEducator project (Contract Nr: ECP 2008 EDU 418006 mEducator) under the eContentplus programme of the European Commission.

8 References

[1] Bizer, C., T. Heath, et al. (2009). Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems (IJSWIS).

[2] IEEE Learning Technology Standards Committee (LTSC). Learning object metadata. In IEEE Learning Technologies Standards Committee 2011. IEEE, 2011.

[3] Dublin core metadata element set, version 1.1, 2011.

[4] MedBiquitous Consortium. Healthcare lom overview. In MedBiquitous Consortium 2011, 2011.

[5] Mitsopoulou Evangelia, Davide Taibi, Daniela Giordano, Stefan Dietze, Hong Qing Yu, Panagiotis Bamidis, Charalampos Bratsas, and Luke Woodham. Connecting medical educational resources to the linked data cloud: the meducator rdf schema, store and api. 1st International Workshop on eLearning Approaches for the Linked Data Age (LinkedLearning2011), 2011.

[6] Pedrinaci, C., Liu, D., Maleshkova, M., Lambert, D., Kopecky, J., and Domingue, J. (2010) iServe: a Linked Services Publishing Platform, Workshop: Ontology Repositories and Editors for the Semantic Web at 7th Extended Semantic Web Conference.

[7] Cohen, T., Schvanevendt, R., , Widdows, D. Reflective Random Indexing and indirect inference: a scalble method for discovery of implicit connections. J Biomed Inform. 43(2):240-56. 2010

[8] Dietze, S., Yu, H.Q., Pedrinaci, C., Liu, D. and Domingue, J. (2011) SmartLink: a Web-based editor and search environment for Linked Services, 8th Extended Semantic Web Conference (ESWC), Heraklion, Greece

[9] Hadzic, M., D'Souza, R., Hadzic, F., Dillon, T. Thinking PubMed: an Innovative System for Mental Health Domain. Computer-Based Medical Systems, 21st IEEE International Symposium on Computer-Based Medical Systems, 2008

[10] Henze, N. (2006). Personalized E-Learning in the Semantic Web. Extended version of 4. *International Journal of Emerging Technologies in Learning (iJET)*, 1(1).

[11] Henze, N., Dolog, P., and Nejdl, W. (2004). Reasoning and Ontologies for Personalized E-Learning. *Educational Technology & Society*, 7(4).

[12] IEEE, IEEE Standard for Learning Object Metadata, *IEEE Std 1484.12.1-2002* , vol., no., pp.i-32, 2002, doi: 10.1109/IEEESTD.2002.94128.

[13] Kaldoudi E, Dovrolis N, Giordano D, Dietze S., Educational Resources as Social Objects in Semantic Social Networks, Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, Herakleio, May 2011

[14] Kobilarov, G. and Dickinson, I. (2008). Humboldt: Exploring Linked Data, Proc. Linked Data on the Web (LDW'08), Beijing, China, April 2008

[15] Li, N., Pedrinaci, C., Maleshkova, M., Kopecky, J. and Domingue, J. (2011) OmniVoke:A Framework for Automating the Invocation of Web APIs, Fifth IEEE International Conference on Semantic Computing, Stanford University, Palo Alto, CA, USA.

[16] Marchionini, G. (2006) Exploratory search: from finding to understanding. Commun. ACM, 49(4):41-46,

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

[18] World Wide Web Consortium, W3C Recommendation, SPARQL query language for RDF, 2008, (http://www.w3.org/TR/rdf-sparql-query/).

[19] Yu, H. Q., Dietze, S., Li, N., Pedrinaci, C., Taibi, D., Dovrolis, N., Stefanut, T., Kaldoudi, E., Domingue, J., A Linked Data-driven & Service-oriented Architecture for Sharing Educational Resources, in Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, CEUR-Vol 717, 2011.

[20] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovrolis, N., Taibi, D., Linked Education: interlinking educational Resources and the Web of Data, in Proceedings of the 27th ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications, Riva del Garda (Trento), Italy, 2012.