A Linked Dataset of medical educational resources

Stefan Dietze, Davide Taibi, Hong Qing Yu and Nikolas Dovrolis

Stefan Dietze is a research group leader at the L3S Research Center (Germany) which he joined in 2011 following previous positions at the Knowledge Media Institute of The Open University (UK) and the Fraunhofer Institute for Software and Systems Engineering (Germany). His research interests are in the areas of Knowledge and Data Engineering, Semantic Web and Linked Data, in particular to address Web data integration problems in actual application domains. Stefan currently is co-ordinator of two European R&D projects, LinkedUp (http://linkedup-project.eu) and DURAARK (http://duraark.eu), both dealing with advancing Linked Data technologies and their take-up in real-world application settings. He also is involved in a number of other projects and initiatives in the Semantic Web area and is co-chair of the KEYSTONE (http://www.keystone-cost.eu/) working group on "Representation of Structured Data Sources" and the W3C Community Group on Open Linked Education. His work has been published in numerous major conferences and journals, he is member of many organization and program committees and editorial boards and a frequent invited speaker. Davide Taibi graduated at the University of Palermo, and currently is a researcher of the Institute for Educational Technologies at the National Research Council of Italy. His research activities are mainly focused on the application of innovative technologies in the e-learning field, with particular emphasis on Mobile Learning, Semantic Web and Linked Data for e-learning, standards for educational processes design, Open Educational Resources. In the last few years, his research has been addressed toward the Learning Analytics field. He has worked as a contract professor at the University of Palermo. Hong Qing Yu is currently works as a lecturer in the Department of Computer Science and Technology at University of Bedfordshire. He is also a key research member of Centre for Computer Graphics and Visualisation (CCGV). He holds a PhD in Computer Science and an MSc in Software Engineering for the e-Economy. He has been involved in 10 research projects. Most of his research projects are funded by European Commissions in the areas of e-learning and e-healthcare systems. He involved in mEducator research project during his post-doctor work in Knowledge Media Institute at The Open University. He is an expert in Semantic Web, Data analytics and Web services research. Nikolas Dovrolis is a Computer Science graduate currently attending an MSc program on "Translational research in Molecular Biology and Genetics." His work delves into biomedical technology and bioinformatics applications. Specifically, his research is focused on Systems biology and Clinical Pharmacology for the School of Medicine, Democritus University of Thrace. Formerly, Nikolas has worked as an Associate Researcher at the same institute developing and maintaining home telehealth applications and e-learning solutions while researching in the areas of social networking, data representation, semantic annotation and services for web applications in telehealth and medical education. He has participated in a number of competitive R&D projects and co-authored journal and proceedings papers in his field of research. Address for correspondence: Dr Stefan Dietze, L3S Research Center, Leibniz University, Appelstraße 9a, 30167 Hannover, Hannover, Germany. Email: dietze@l3s.de

Abstract

Reusable educational resources became increasingly important for enhancing learning and teaching experiences, particularly in the medical domain where resources are particularly expensive to produce. While interoperability across educational resources metadata repositories is yet limited to the heterogeneity of metadata standards and interface mechanisms with a lack of shared or aligned controlled vocabularies, Linked Data (LD) principles, based on W3C standards and supported through a wide range of tools, open up opportunities to alleviate such problems. We introduce the "mEducator Linked Educational Resources" dataset, which offers a range of open educational resources for the medical domain, exposed through LD principles. Data have been generated through a combination of manual curation and semi-automated harvesting

techniques, and state-of-the-art enrichment and clustering techniques were deployed in order to classify and categorize data, toward improved reusability and access. Data are currently used by a range of educational applications and is accessible for third parties and developers, for instance through the LinkedUp Catalog and other registries, to facilitate further take-up and applications.

The dataset

- Location
  - Dataset described at http://linkededucation.org/meducator and http://datahub.io/dataset/meducator
  - SPARQL endpoint: http://meducator.open.ac.uk/resourcesrestapi/rest/meducator/sparql
  - Dump: http://lak.linkededucation.org/meducator/meducatorDump.rdf
- Creator: Stefan Dietze, Hong Qing Yu, Davide Taibi
- Date: released 06/2010 (continuously updated)
- Format: application/rdf+xml
- Restrictions to use: Creative Commons – BY license

Introduction

Sharing and reusing educational resources have long been an overall vision within the open educational resources (OER) and technology-enhanced learning (TEL) communities. Particularly in the medical domain where high-quality educational resources are particularly expensive to produce, reuse of existing reuse, within and across organizational boundaries, is an important aim. While the TEL community has provided a range of techniques, metadata standards and interface mechanisms (Dietze & Sanchez-Alonso *et al*, 2013), interoperability is still hindered by the heterogeneity of approaches and the lack of controlled, shared or aligned vocabularies. This has led to vast amounts of educational resource metadata becoming available on the Web as part of still rather isolated and disparate educational resource metadata silos, such as ARIADNE (http://www.ariadne-eu.org/), OpenLearn (http://www.open.edu/openlearn/) or the different repositories of the OpenCourseWare Consortium.

Here, Linked Data (LD) principle (Bizer, Heath & Berners-Lee, 2009), based on W3C standards and supported through a wide range of tools, emerged as de facto standard for sharing data on the Web and opens up opportunities to alleviate such problems. We introduce the "mEducator Linked Educational Resources" dataset, which offers a range of OER for the medical domain, exposed through LD principles. Since a large amount of educational data is already available on the Web via proprietary and/or competing schemas and interface mechanisms, the main challenges addressed by the mEducator dataset are to (1) start adopting LD principles and vocabularies while (2) leveraging on existing educational Web data accessible via non-LD compliant means and (3) improving interoperability at different levels, such as interfaces and schemas. The dataset, currently consisting of 780 educational resources, has been generated through a combination of manual curation and semi-automated harvesting techniques following two distinct mechanisms: (1) data are added and curated manually by users of the educational social network *MetaMorphosis* (Dietze, Kaldoudi, Dovrolis, Yu & Taibi, 2011) and (2) metadata is extracted from 10 heterogeneous educational repositories by deploying mechanisms for data lifting from heterogeneous schemas and formats into a unified RDF schema. State-of-the-art enrichment and clustering techniques were deployed in order to classify and categorize data, toward improved reusability and access. The dataset is one of the central outcomes of the mEducator project (Dietze *et al*, 2011).

Table 1: Resource types and properties and their population in the mEducator dataset

Resource type	#	*Properties*	*%*
mdc : Resource	780	mdc : title	98.5
		mdc : description	97.3
		mdc : subject	97.3
		mdc : creator	97.3
		mdc : language	89.6
		mdc : rights	75.3
		mdc : educationalContext	63.8
		mdc : educationalLevel	62.9
		mdc : teachingLearningInstructions	56.9
		mdc : educationalObjectives	56.5
		mdc : educationalOutcomes	50.8
		mdc : educationalPrerequisites	49.2
		mdc : disciplineSpeciality	45.8
		mdc : assessmentMethods	39.1
		mdc : hasEnrichmentContext	36.4
mdc : EnrichmentContext	1351	mdc : hasEnrichment	100
		mdc : enrichmentType	100
foaf : Person	1362	foaf : name	100
		mdc : profileURI	51,1
		mdc : title	1
mdc : Algorithm	12532	mdc : hasConfidenceLevel	100
		mdc : algorithmName	100
		mdc : algorithmDescription	100
mdc : Cluster	2722	mdc : hasCommonFeatures	100
		mdc : containsResource	100
		mdc : associatedTo	2

The dataset is currently used by a range of educational applications and is accessible for third parties and developers, for instance through the LinkedUp catalog and other registries, to facilitate further take-up and applications.

Creating Linked Open Data of medical educational resources

Based on studies of existing vocabularies and schemas in the educational field, the mEducator schema (Mitsopoulou *et al*, 2011) was generated, being one of the first native RDF schemas for educational resources, fundamentally based on existing vocabularies and schemas and linked to a number controlled vocabularies for description of subject and licensing features. The schema (available at http://www.purl.org/meducator/ns/) covers the most frequently used aspects of educational resources—from basic ones such as title and descriptions to more sophisticated ones such as learning outcomes and licensing models. Data have been generated by extracting data from existing datasets and repositories (see Dietze *et al*, 2011 for details) and by manually annotating and curating extracted metadata annotations. All generated educational metadata is eventually stored in a dedicated RDF store as part of the mEducator dataset (more details at http://datahub.io/dataset/meducator). Each educational resource owns a unique and de-referencable URI (see for an example http://purl.org/meducator/resources/25a8c581-66d7-4186-9411-f9f0f783463e). A SPARQL endpoint is available at http://meducator.open.ac.uk/resourcesrestapi/rest/meducator/sparql.

Table 1 provides an overview of the total number of resources per type, their associated properties and the percentage of their population, ie, the proportion of resources of the respective type showing a description of the respective property.

For improved data access and to facilitate data reuse, 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 access updating related APIs). These REST APIs facilitate different types of queries, such as generic SPARQL queries, keyword-based queries, property-based keyword queries, *rdfs:seeAlso*-based keyword queries or an identifier-based property queries. The proposed architecture has been introduced in (Dietze *et al*, 2012) and is fully described in Dietze and Kaldoudi *et al* (2013). While at the moment no major work and expansion of the dataset are carried out, the data are maintained continuously, and further expansion is envisaged in future work.

Data enrichment, clustering and interlinking

Though metadata is lifted into RDF automatically, it is often poorly structured and makes only very limited and fragmented use of controlled vocabularies. To alleviate this problem, we have further enriched the resource metadata by taking advantage of available datasets and vocabularies such as DBpedia (via DBpedia Spotlight) and the multitude of ontologies available via the BioPortal APIs. These enrichments are utilized for (1) expanding existing metadata with publicly available knowledge, (2) disambiguation of data and (3) clustering correlated resources by exploiting the use of shared vocabularies.

Data enrichment is implemented in two ways (1) as an automated mechanism whenever new data is pushed to the RDF store and (2) as semi-automated approach that provides end users with suggestions of related entities that match a particular term, implemented as a feature into the *MetaMorphosis* application (Dietze *et al*, 2011). While the first approach makes usage of DBpedia exclusively, resulting in large numbers of automatically retrieved references to DBpedia resources, the second approach makes exclusive use of the BioPortal API which provides access to over 300 biomedical ontologies and 5 million entities. Figure 1 shows a subgraph of our data graph where green circles denote DBpedia entities and blue squares represent educational resources. A particular cluster of resources sharing enrichments indicating some shared subjects in the cardiological domain is highlighted.

The enrichments were evaluated in earlier work (Dietze and Kaldoudi *et al*, 2013), showing a 92% precision for retrieved annotations. In addition, a range of clustering techniques was applied in order to identify significant clusters forming a subject-related set of resources.

Descriptive statistics

The mEducator Linked Educational Resources RDF repository contains in total 292 258 triples, of which 33 011 directly referred to a total 780 distinct educational resources. The average number of triples per educational resource is 27, ranging from a minimum of six to a maximum of 68. Figure 2 provides an overview of the heterogeneity of the origin of individual resource metadata.

In addition, the metadata imported from external stores is usually very limited, and it often covers only few properties, eg, title, description and resource location. Based on our automated and semi-automated enrichment techniques, richer descriptions that populate additional properties are provided for the majority of resources. As a result, all resources have a minimum of five described properties. The usage frequency of particular properties in our RDF schema reflects the importance of particular schema elements (eg, when considering the source metadata descriptions) and the richness of the transformed and enriched resources descriptions (when considering the improved metadata in our dataset). A more detailed description of individual properties is given in Mitsopoulou *et al* (2011).

Applications, impact and usage

Soon after the first release of the mEducator dataset, four educational Web applications have been developed that interact with the mEducator dataset. For example, the data and services

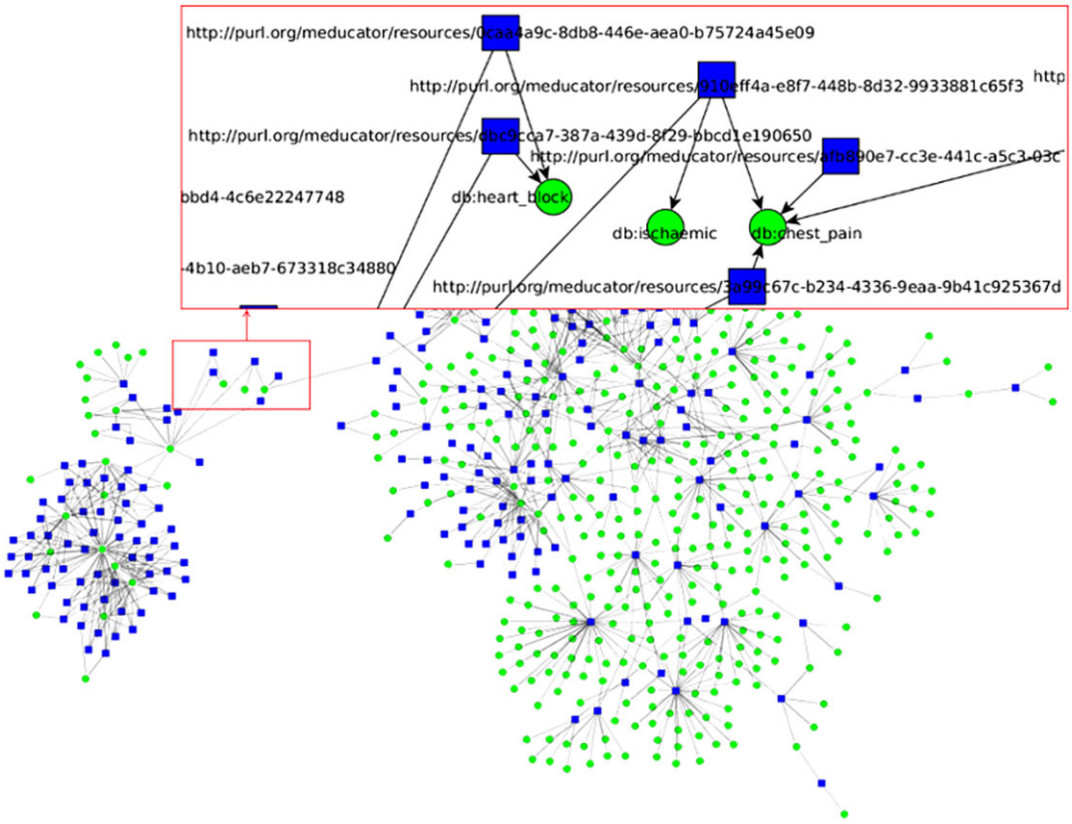

Figure 1: Relationships between educational resources (blue, square nodes) from distinct repositories based on shared DBpedia references (green, circular nodes), indicating similar learning subjects in the field of cardiology

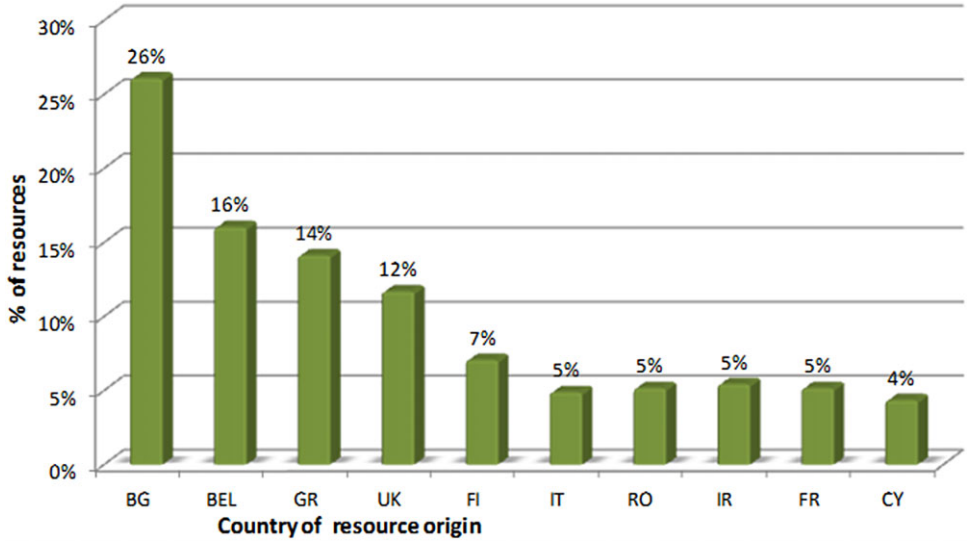

Figure 2: Number of resources (%) per country of origin (based on country of origin of contributing institution)

integration APIs and datasets presented in the previous sections are fully integrated in the *MetaMorphosis* (Dietze *et al*, 2011) environment, which merges the paradigms of semantic and social web for sharing linked educational resources. *MetaMorphosis* realizes the educational application and presentation layer. It allows viewing, management and annotation of the educational resource metadata that is retrieved via the APIs provided by the educational services and data integration mechanisms introduced here. Meanwhile, additional medical educational applications have emerged which interact with our mEducator dataset, which are *Miles+* (available at http://kedip.med.auth.gr/meducator3/milesplus/), *Melina+* (available at http://www.meducator3.net/melinaplus/) and *Linked Labyrinth* (available at http://www.meducator3.net/llplus/) and take advantage of enrichments and clustering to provide precise recommendation and search results.

The impact of the dataset, however, reaches beyond individual applications. Though being a comparably small dataset, it has served as an initial proof-of-concept which helped to establish a set of principles for educational resource metadata sharing, further described in (Dietze *et al*, 2012). At the time of initiating the mEducator dataset, LD principles had not been widely reflected and applied in the TEL community despite its strive for sharing and reuse of educational resources and data. The mEducator dataset has been one of the first adaptors and positive examples for sharing OER which paved the way for a constantly growing community in the area. Initiatives such as LinkedEducation (http://linkededucation.org), LinkedUniversities (http://linkeduniversities.org), the LinkedUp project (http://linkedup-project.eu) or the more recent W3C Community Group on Linked Open Education (http://www.w3.org/community/opened) now all embrace such principles in the educational domain, in parts being inspired by the work described here. This ever growing community has led to the emergence of a wide variety of educational LD available on the Web, with the LinkedUp Data Catalog (http://data.linkededucation.org/linkedup/catalog/) being one of the largest registries, also including the mEducator dataset.

Ethical considerations

Given the nature of the dataset—containing metadata of educational resources and no personal data—privacy and data protection issues are no concern. All data are publicly available under open licenses and either has been natively added to our dataset by users or was harvested using publicly available APIs. No actual user or medical data are captured in any way, beyond simple attributions to the resource authors, manually added by the user themselves through a simple reference to their name. No other personal data such as address or contact details are captured in our dataset. Additional personal information is captured through the mentioned applications such as *MetaMorphosis*; however, such data are out of scope of our dataset and are kept separately from our resource metadata.

Limitations

Given the semi-automated extraction and curation of data, our dataset is currently including a limited number of resources. Future work will focus on two major areas, including the investigation of methods to enable the integration of data from other educational domains, and the extension of the framework with additional open repositories and data stores. In addition, as shown by our dataset analysis, only a limited set of properties is actually populated by the actual annotators, either directly in our dataset or within the remote repositories from which data have been extracted, where additional semantics are added through our data enrichment techniques.

Acknowledgements

This work has been partly funded by the mEducator project (Contract number: ECP 2008 EDU 418006 mEducator) under the eContentplus program of the European Commission.

References

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.

Dietze, S., Kaldoudi, E., Dovrolis, N., Yu, H. Q. & Taibi, D. (2011) *MetaMorphosis*+—a social network of educational Web resources based on semantic integration of services and data. 10th International Semantic Web Conference (ISWC2011), Bonn, Germany.

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, Riva del Garda (Trento), Italy, 2012.

Dietze, S., Sanchez-Alonso, S., Ebner, H., Yu, H. Q., Giordano, D., Marenzi, I. *et al* (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.

Dietze, S., Kaldoudi, E., Dovrolis, E., Giordano, D., Spampinato, C., Hendrix, M. *et al* (2013). Socio-semantic integration of educational resources—the Case of the mEducator Project. *Journal of Universal Computer Science (J.UCS), 19*, 11, 1543–1569.

Mitsopoulou, E., Taibi, D., Giordano, D., Dietze, S., Yu, H. Q., Bamidis, P. *et al* (2011). *Connecting medical educational resources to the Linked Data cloud: the mEducator RDF Schema, Store and API*. Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, CEUR-Vol 717, 2011.