This document is a preliminary version of the book chapter:

D. Giordano, S. Dietze, C. Spampinato, D. Taibi, E. Kaldoudi, N. Dovrolis, E. Mitsopoulou, H.Q. Yu, S. Konstantinidis, B. Charalampos, P. Bamidis (2011). Towards linking educational resources on the web through clustering and enrichment: the mEducator schema. (pp. 121-133) In: L. Despotova-Toleva, V. Anastasov, P. Bamidis (Editors) E-education & E-science. Medical Publishing VAP; Plovdiv, Bulgaria: 2011 (ISBN 978-960-243-682-0)

Towards linking educational resources on the web through clustering and enrichment: the mEducator schema

Daniela Giordano[1] (dgiordan@dieei.unict.it)

Stefan Dietze[2] (dietze@l3s.de)

Concetto Spampinato[1] (cspampin@dieei.unict.it)

Davide Taibi[3] (davide.taibi@itd.cnr.it)

Eleni Kaldoudi[4] (kaldoudi@med.duth.gr)

Nikolas Dovrolis[4] (dovroli@alex.duth.gr)

Evangelia Mitsopoulou[5] (emitsopo@sgul.ac.uk)

Hong Qing Yu[6] (h.q.yu@open.ac.uk

Stathis Konstantinidis[7] (cs@med.auth.gr

Bratsas Charalampos [8] (cbratsas@auth.gr)

Panos Bamidis[7] (bamidis@med.auth.gr

[1]Dept. of Computer Engineering, University of Catania, Italy
[2]L3S research Center. Leibniz University, Hanover, Germany
[3]Educational Technology Institute CNR, Palermo, Italy
[4]Democritus University of Thrace Alexandroupolis, Greece
[5]Center for Medical Education, St.George University of London,UK
[6]Knowledge Media Institute, Open University, Milton Keynes, UK
[7]Medical Informatics Lab, Aristotle University of Thessaloniki, Greece
[8]Mathematical Dept, Aristotle University of Thessaloniki, Greece

Abstract
A crucial requirement in the web of data is to describe datasets to facilitate access, reuse, and interlinking. The mEducator project, a best practice network on medical content sharing and repurposing has developed an RDF schema that departures from current learning metadata standards by adopting linked data principles. We describe the design rationale of the schema, and focus on two innovations, i.e., the "clustering" and "enrichment" classes, that, for each educational resource, store explicitly connections with related resources (computed by clustering algorithms) and with semantically related concepts, respectively. The schema has been implemented in mEducator content sharing solutions and we elaborate on its potential benefits and applications, both in the e-learning and in the e-science fields.

Author Keywords
metadata; linked data; educational resources; clustering; enrichment; standards; semantic Web

General Terms
Standardization, design, documentation, content sharing, e-science

Introduction

A crucial requirement of the modern web of data is to describe datasets to facilitate access, reuse, and interlinking, as best exemplified by the LOD (Linked Open Data) cloud project and by the linked data principles to publish datasets in the web [3]. The mEducator Best Practice Network (www.meducator.net) is devoted to the development and critical assessment of solutions for multi-type medical content sharing and repurposing based on web 2.0 and web 3.0 technologies, and has developed an RDF metadata schema that departures from current e-learning metadata standards by 1) adopting linked data principles [3] and 2) by introducing some innovations that should support novel ways of sharing and accessing resources, overcoming through interlinking [4] the problem of isolated data "silos". The mEducator schema has been designed to support annotation from end users through either institutional content management systems or portals that are a centralized point of access to distributed mEducator repositories and utilize the underlying mEducator infrastructure. The challenges that mEducator has addressed are, on the one hand, to provide a standardized annotation mechanism for educational resources that is easy to use [1] but powerful enough to support and track repurposing activites, and, on the other hand, to support flexible and dynamic ways to explore distributed datasets.

The mEducator RDF Schema: overview

The purpose of the mEducator"s metadata scheme [15] is to provide a standardised format to describe medical educational resources in order to share, search, retrieve and repurpose them across academic institutions. Resource Description Framework (RDF) and its serialisation in XML were identified as the most appropriate data model and representation languaguage to implement the metadata in a standardised format. The rationale behind this choice was to be compliant with the Linked Data approach to publish data on the Web, and to leverage on the power of semantic web technologies to interpret and merge information from multiple sources. In this way, the mEducator resources can be "understood" by third party applications, and, on their turn, can benefit from the "knowledge" available from other datasets, vocabularies, ontologies to improve the quality of the mEducator resources's descriptions. The design philosophy of the schema was to keep it as lightweight as possible and to maximise its interoperability and reusability.

The IEEE Learning Object Metadata (LOM) [10] and its extension Healthcare LOM, which is currently used by Medbiquitous (http://www.medbiq.org/) for the description of educational resoures in the healthcare field, were not adopted by mEducator, since the serialisations of these metadata schemas have been officially expressed directly in XML. Instead we focused on Dublin Core[1], which is has been officially expressed in RDF/XML. We reused and customized Dublin Core elements, following the current trend in metadata design, i.e. the mix and match of vocabularies from other well established schemas. In turn, the novel properties/vocabularies introduced in the mEducator schema were designed to form independent, reusable structures that possibly would be meaningful also for other communities. A few mEducator properties were defined as sub properties of Dublin Core ones and mEducator specific controlled vocabularies were developed for the properties Resource type, Media type, Educational level, Educational outcomes. These vocabularies were designed to easily separate mEducator peculiar aspects (i.e., specific to the medical/healthcare field) from the more general ones [8]. The serialisation of controlled vocabularies in RDF via SKOS[2] replaced the free text values in some properties of the schema with SKOS URI's. The personal profiles of the resource's creator and of the metadata creator are described via FOAF[3]. Table 1 on the right lists the main schema properties. Among these, some are devoted to track any repurposing of the resource, i.e. the modifications that have been performed to make the resource fit to a different context (e.g. a different language, different target audience, different educational goal, different cultures, these latter w.r.t. to metric system in use, procedures, sensitive issues, behavioral habits). Other properties, namely, mdc:rights and mdc:quality are devoted to express how any resource is licensed w.r.t. Intellectual Property Rights, and therefore under what terms it can be reused or repurposed; and

[1] http://dublincore.org/documents/dces/
[2] www.w3.org/2004/02/skos/
[3] www.foaf-project.org/

if any quality stamp or certification has been awarded to the resource, or if it has undergone some official review process. Along the course of the project we have realized the need to extend our schema with two novel classes, Enrichment and Clustering. These classes address, respectively: 1) the need to improve the quality of the resource's annotation when no controlled vocabulary is used, and 2) the need to have an additional support for the "exploratory search" of resources from distributed sources, as opposed to "focused search" with a specific, well-defined search goal [7].

Property	Range	Mul
mdc:assessment Methods	String	0.. ∞
mdc:citation	String	0..1
mdc:created	dc:date	0..1
mdc:creator	foaf:Person	0.. ∞
mdc:description	String	0..1
mdc:discipline	String	0.. ∞
mdc:educational Context	String	0.. ∞
mdc:educational Level	String	0.. ∞
mdc:educational Objectives	String	0.. ∞
mdc:educational Outcomes	skos: Concept	0.. ∞
mdc:educational Prerequisites	String	0.. ∞
mdc:identifier	Literal	1.. ∞
mdc: isRepurposedFrom	mdc: Resource	0.. ∞
mdc:language	ISO: Language	1.. ∞
mdc:mediaType	skos: Concept	0.. ∞
mdc: metadataCreated	ISO: Language	1
mdc: metadataCreator	foaf:Person	1.. ∞
mdc:metadata Language	ISO: Language	1.. ∞
mdc:quality	String	0.. ∞
mdc:repurposing Context	skos: Concept	1.. ∞
mdc:repurposing Description	String	0.. ∞
mdc:resourceType	skos: Concept	0.. ∞
mdc:rights	String	1
mdc:subject	Literal	1.. ∞
mdc: technicalDescription	String	0.. ∞
mdc:title	String	1

Table 2. Main properties (with Range and Multiplicity) of the mdc:Resource (i.e. mdc:Resource is the domain of the above properties). The full RDF schema can be found at http://purl.org/meducator/ns

The Enrichment class

Educational metadata is often poorly structured, as it is based on either unstructured text or less than well-defined terminologies. Thus, such metadata needs to be enriched. In particular, the Linked Data cloud already offers large amounts of datasets, ranging from general-purpose ones like DBpedia to domain-specific educational datasets. To this end, we have developed enrichment mechanisms which automatically enrich poorly structured descriptions with links to related terms in well-established vocabularies. To this end, the RDF schema was expanded with two additional concepts: <mdc:EnrichmentContext> and <mdc:Enrichment>. Each resource can be associated with multiple enrichment contexts which in turn refer to exactly one particular mdc:Enrichment

instance. The latter describes the actual enrichment, e.g. a reference to a particular DBpedia resource (such as http://dbpedia.org/resource/Polymerase in Figure 1. The enrichment context is used to further describe the relationship between the resource and the enrichment. This can include, for instance, the property of the resource the enrichment relates to (e.g. mdc:title or mdc:subject) or the confidence level of the enrichment. Avoiding a direct association of resources with enrichments also helps in avoiding multiple duplications of the same enrichments: in a large dataset it is expected that many mdc:Resources relate to the same enrichment, though the type and coherence of each relationship might differ. While enrichment of unstructured data poses several crucial research challenges such as entity recognition and text mining, we take advantage of available and established APIs such as the ones provided by *DBpedia Spotlight*[4] and *BioPortal*[5], which partially tackle some of the related issues and allow access to a vast number of established taxonomies and vocabularies, such as *DBpedia*, *SNOMED*[6], *MESH*[7] or *Galen*[8]. That way, unstructured free text, for instance the keyword "thrombolysis", is enriched with unique URIs of structured LD entities - such as *http://dbpedia.org/resourceThrombolysis* which refers to the corresponding DBpedia resource or *http://www.co-ode.org/ontologies/galen#Thrombolysis* referencing to a matching concept within the GALEN ontology. Enrichments allow further reasoning on related concepts and also enable users to query for resources by using well-defined concepts and terms as opposed to ambiguous free text. Figure 1 depicts an example RDF resource description after enrichment. In particular, the listing shows a reference from a mdc:Resource to a particular mdc:EnrichmentContext, which in turn links the resource to a particular enrichment by allowing the further description of the enrichment, such as the enrichment type. Enrichment is implemented currently in two different ways: (a) as automated mechanism whenever new data is pushed to the RDF store where all the resources retrieved by distributed queries are cached for better processing; 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 [5]. While the first approach makes use of DBpedia exclusively, resulting in large numbers of automatically retrieved referencesto DBpedia resources, the second approach makes use of the BioPortal API exclusively.

```
    <mdc:hasEnrichmentContext rdf:resource="http://meducator.open.ac.uk/ontology/context
    /9483973089370398709818ß09fkjkdu9084"/>
  </mdc:Resource>
- <mdc:Enrichment rdf:about="http://meducator.open.ac.uk/ontology/dbpedia.org/resource/Polymerase">
    <rdfs:isDefinedBy rdf:resource="http://dbpedia.org/resource/Polymerase "/>
    <rdfs:label>Polymerase</rdfs:label>
    <mdc:externalSource>DBPedia</mdc:externalSource>
  </mdc:Enrichment>
- <mdc:EnrichmentContext rdf:about="http://meducator.open.ac.uk/ontology/context/9483973089370398709818ß09fkjkdu9084">
  - <rdfs:comment>
      The URI should be generated automatically by SESAME on upload.
    </rdfs:comment>
    <mdc:enrichmentType>Property reference: mdc:title</mdc:enrichmentType>
    <mdc:hasEnrichment rdf:resource="http://meducator.open.ac.uk/ontology/dbpedia.org/resource/Polymerase"/>
  </mdc:EnrichmentContext>
</rdf:RDF>
```

Figure 1: An instance of the mEducator Enrichment class

[4] http://dbpedia.org/spotlight
[5] http://www.bioontology.org/wiki/index.php/BioPortal_REST_services
[6] http://www.ihtsdo.org/snomed-ct/
[7] http://www.nlm.nih.gov/mesh/
[8] http://www.co-ode.org/galen/

The Clustering class

In the logic of the Semantic Web, better search and navigation is achieved through semantic links [12]. In the semantic search paradigm, as in conventional web search paradigms, it is assumed that the user is capable to issue a query that precisely reflects her information needs. However, often this is not the case, because of lack of knowledge about the vocabulary used for indexing, or because users engage in an "exploratory search" mode to focus their informational need, especially when the goal is learning about complex and/or unknown topics [13]. Clustering refers to machine learning techniques that group resources based on their similarities [17]; from an information retrieval perspective, clustering improves the recall of potentially relevant resources and may provide the user with hints through inspection of the cluster contents [6]. To support not only exploratory search of the content of local repositories, but also the interlinking of otherwise distributed web resources [4], we have defined a new RDF schema to provide information on clusters of similar resources. Our clustering schema (Figure 2) adopts the Provenance Vocabulary[9] and contains a set of classes and properties to describe a complete clustering phase: from the content of each cluster, to the employed algorithm, to the distinctive features of each cluster, etc.

In detail, the *Cluster* class is *SubClassOf* a *mdc:Resource* since a cluster can be seen itself as a mEducator resource. Each cluster may contain a set of *mcc:ClusterDataItem,* subclass of *mdc:Resource* which in turn is subclass of *rdf:resource.* This allows us to cluster, and consequently to link, not only mEducator resources but also external resources. Each cluster can be also associated to another cluster (*associatedTo*) thus introducing the concept of similarity among clusters (and not only among resources). This notion may be employed to increase the depth of the exploratory search process, to pass from exploration of one cluster to exploration of the more closely related clusters. Each cluster is also characterized by a set of discriminative features (*hasCommonFeatures*), i.e. the features in common between the resources of the considered cluster. The schema, finally, contains information (*Actor, date* and *Algorithm*) about the execution process (*DataCreation*) to obtain the clusters.

For each cluster we track the software agent who executed the clustering process, the date when the clustering was performed and the employed clustering algorithm together with its description, confidence level and used features. This turns out to be important, since, as pointed out in [9], research on clustering of semantic web resources is still in its beginnings, and the performance of any given clustering method is deeply affected by the nature of the data set to be clustered.

[9] http://trdf.sourceforge.net/provenance/ns.html

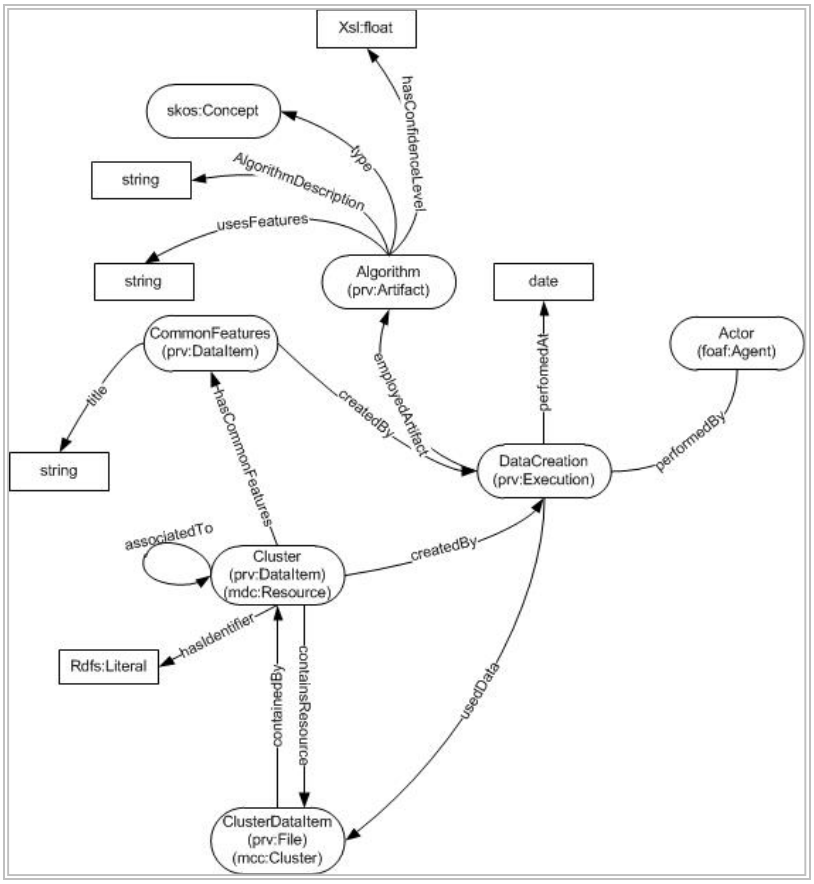

Figure 2. Clustering Schema. *prv* is the namespace for the Provenance vocabulary

First experiences and future scenarios

The mEducator schema has been implemented in all the mEducator content sharing solutions. These include the MetaMorphosis+ semantic social network [14], the MILES+ environment [2], the Melina+ environment (http://www.meducator3.net/drupal) and the mash-up based mEducator 2.0 (www.meducator2.net). All these utilize the RDF/xml instances produced by the schema for both import and export of the metadata.

For the actual enrichment, a special user interface was developed [5], to facilitate the users in filling the metadata fields keyword, discipline and specialty (within that discipline) with terms enriched via the BioPortal ontologies, as described in the previous sections. All the user is asked to do is to launch a search for the term and then decide which ontology he prefers to use, among the ones retrieved from BioPortal and proposed. This approach enables us to annotate the user's terms linking them to specific ontologies and thus gathering more accurate and "contextual" information utilizing the user's knowledge.

A common finding across the implementations was the complexity and size of the input required by the user. The forms tend to become cluttered and counter-intuitive. After some experimentation with different approaches, better reactions from the users were achieved when the schema fields were organized in groups and when a "hint" system was provided. In general, concepts like multiplicity and requirement could be made more apparent to the end user with the use of the appropriate interface elements. Modern web design technologies like ajax, javascript, flash or html5 coupled with libraries like jquery can help create a better result, although this presents some challenges to the developers since creating the RDF from text boxes and dropdown menus in a user interface is a complex matter.

Linking resources to shared LOD vocabularies serves two main purposes. Firstly, it allows expanding the metadata of individual resources with additional knowledge. Secondly, it provides a means to identify correlations between individual resources which share the same external references. Hence, it also offers opportunities to cluster related resources. On the other hand, the use of machine learning techniques to cluster resources as proposed in [7] affords a complementary advantage, since similarites are detected based on methods do not rely on named entity recognition, and therefore can catch more subtle linguistic patterns. The exploratory search has been already prototyped in [14], and currently we are engineering a configurable clustering Web service that can operate on any RDF repository and generate output compliant with the presented clustering schema. The results of the clustering, whenever they highlight stable similarity relationships across resources from distributed repositories, offer also a valuable source to assess in order to automaticallay interlink such repositories.

The enrichment and clustering notions, introduced and formally modeled within the mEducator metadata schema can be considered first implementations of a modern approach to describing educational resources, quite generalizable to learning domains other than the medical one, and more in line with the trends in place in the e-learning and the e-science communities. There are some interesting parallels that can be drawn between the e-learning and the e-science needs with respect to metadata. The tracking of the repurposing of learning resources is similar to the need of tracking the provenance of the scientific datasets and of the computations that have been perfomed to perform a scientific investigation; the need to support discovery and reuse of datasets/educational resources across different domains and communities is shared in both

fields. In e-science the same approach of annotating any data set through LOD enrichments to achieve more accurate descriptions, and performing some clustering of the metadata descriptions to highlight hidden patterns across such descriptions and can serve the same purpose as in e-learning: affording better chances of serendipitous discovery of relevant items. The two proposed metadata classes, designed and described according to semantic web standards, offer a promising, concrete tool to achieve the vision of better and more robust interlinking of distributed resources in the Web.

Acknowledgements

This work was supported by the project mEducator (Multi-type Content Sharing and Repurposing in Medical Education), funded by the eContentplus Programme, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable (Grant ECP 2008 EDU 418006).

We also thank all the members of the mEducator TRG (Technical Reference Group) who helped shaping the schema.

References

[1] Bamidis PD, Nikolaidou M, Konstantinidis S, Kontakiotis T, Triaridis S, Giordano D, and Davies P. "The burden of completing educational metadata for digital resources: a focus group study on user perceptions", In *Proc. of the MedBiquitous Annual Conference 2011*

[2] Bamidis, P., Konstantinidis, S.T., Bratsas, C. and Iyengar, M.S. Federating learning management systems for medical education: A persuasive technologies perspective. In *Proc. CBMS 2011*, IEEE (2011), 1-6.

[3] Bizer, C., Heath, T. and Berners-Lee, T. Linked data - The Story So Far. *Int. J. on Semantic Web and Information Systems (IJSWIS),* 5, 3 (2009) 1-22.

[4] Dietze, S., Yu, H. Q., Giordano, D., Kaldoudi, E., Dovorolis, N., and Taibi, D. Linked education: interlinking educational resources and the web of data. In *Proc. SAC 2012* ACM (2012)

[5] Dovrolis, N., Stefanut, T., Dietze, S., Yu, H. Q., Valentine, C. and Kaldoudi, E. Semantic Annotation and Linking of Medical Educational Resources. In *Proc. 5th European Conf. of the International Federation for Medical and Biological Engineering*, vol. 37, Springer (2011), 1400-1403.

[6] Faro, A., Giordano, D, and Santoro C. Link-based shaping of Hypermedia Webs Assisted by a Neural Agent. *Journal of Universal Computer Science (JUCS)*, 4 7, 1998.

[7] Giordano, D. Faro, A., Maiorana, F., Pino, C. and Spampinato, C. Feeding back learning resources repurposing patterns into the "information loop": opportunities and challenges. In *Proc. ITAB 2009, 9th Int. Conf. on Information Technology and Applications in Biomedicine*, IEEE (2009).

[8] Giordano, D., Kavasidis, I., Spampinato, C. and Bamidis, P. Developing controlled Vocabularies for educational Resources Sharing: a Case Study. In *Proc. Linked Learning 2011: 1st Int. Work. on eLearning Approaches for the Linked Data Age*, CEUR-Vol 717, 2011

[9] Grimnes, G. A., Edwards, P. and Preece, A. (2008). Instance based clustering of semantic web resources. In *Proc. of the 5th European semantic web conference.* LNCS 502, Springer-Verlag (2008) 303-317

[10] 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.

[11] Kaldoudi E, Dovrolis N, Giordano D, Dietze S., Educational Resources as Social Objects in Semantic Social Networks. In *Proc. Linked Learning 2011: 1st Int. Work. on eLearning Approaches for the Linked Data Age*, CEUR-Vol 717, 2011

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

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

[14] Metamorphosis+ Social Network. http://metamorphosis.med.duth.gr

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

[16] 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 *Proc. Linked Learning 2011: 1st Int. Work. on eLearning Approaches for the Linked Data Age*, CEUR-Vol 717, 2011

[17] Xu, R. and Wunsch II, D. C. (2005). Survey of clustering algorithms. Ieee Trans. on Neural Networks, 16 (3), pp 645-678 .