Journal of Universal Computer Science, vol. 19, no. 11 (2013), 1543-1569
submitted: 15/12/12, accepted: 28/5/13, appeared: 1/6/13 © J.UCS

Socio-semantic Integration of Educational Resources - the Case of the mEducator Project

Stefan Dietze
(L3S Research Center, Leibniz University, Hanover, Germany
dietze@l3s.de)

Eleni Kaldoudi, Nikolas Dovrolis
(Democritus University of Thrace, Alexandroupoli, Greece
kaldoudi@med.duth.gr, ndovroli@alex.duth.gr)

Daniela Giordano, Concetto Spampinato
(University of Catania, Dipartimento di Ingegneria Elettrica, Elettronica e Informatica, Italy
{dgiordan, cspampin}@dieei.unict.it)

Maurice Hendrix, Aristidis Protopsaltis
(Serious Games Institute, Coventry University, UK
{MHendrix, AProtopsaltis}@cad.coventry.ac.uk)

Davide Taibi
(Italian National Research Council, Institute for Educational Technologies, Palermo, Italy
davide.taibi@itd.cnr.it)

Hong Qing Yu
(Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
h.q.yu@open.ac.uk)

Abstract: Research in technology-enhanced learning (TEL) throughout the last decade has largely focused on sharing and reusing educational resources and data. This effort has led to a fragmented landscape of competing metadata schemas, such as IEEE LOM or ADL SCORM, and interface mechanisms, such as OAI-PMH, SQI and REST-ful services in general. More recently, semantic technologies were taken into account to improve interoperability. However, so far Web-scale integration of resources is not facilitated, mainly due to the lack of take-up of shared principles, datasets and schemas. On the other hand, the Linked Data approach has emerged as the de facto standard for sharing data on the Web and is fundamentally based on established W3C standards. This paper presents results of the European Commission-funded project mEducator, which exploits Linked Data principles for (1) semantic integration and (2) social interconnecting of educational data, resources and actors. We describe a general approach to exploit the wealth of already existing educational data on the Web by allowing its exposure as Linked Data and by taking into account automated enrichment and interlinking techniques to provide a rich and well-interlinked graph for the educational domain. Additionally, the paper presents an evaluation of our work with respect to a set of socio-semantic dimensions. Experimental results demonstrate improved interoperability and retrievability of the resulting resource descriptions as part of an interlinked resource graph.

Keywords: Linked Data, Semantic Web, SOA, Technology-enhanced Learning, Clustering
Categories: M.1, M.4, M.7, M.8, M.9, H.3, H.4, H.5

1 Introduction

The most recent generation of Web based learning technologies uses the Internet as a means to create active, context-based, personalized learning experiences shifting the emphasis from 'teaching' to 'learning' and from the notion of technology as a didactic mediator to the notion of sociable, peer-supported, involved learner.

Stakeholders in this new manifestation of technology enhanced learning (TEL) point out that access to comprehensive repositories of learning content and metadata is one of the most crucial factors in the future of learning. Efforts in this endeavor have led to a fragmented landscape of (sometimes competing) metadata schemas, such as Dublin Core (http://dublincore.org/documents/dces/), IEEE Learning Object Metadata (LOM) [IEEE, 02] or ADL SCORM (http://www.adlnet.org) and query interface mechanisms such as OAI-PMH (http://www.openarchives.org/OAI/openarchivesprotocol.html) or SQI (http://www.cen-ltso.net/main.aspx?put=859) which are exploited by educational resource repository providers to support data interoperability. Moreover, the current explosion of Linked Data (LD) [Bizer, 09; Heath, 11] repositories is gradually leading to an ever-growing amount of data sets and schemas for different fields that support wide-scale data interoperability based on generic Internet standards. Although not widely adopted by the educational community yet [Dietze, 13], RDF[1]-based datasets are expected to add to this wealth and variability of educational content repositories. Based on these standards, a vast amount of educational content and related metadata are currently shared on the Web. However, while there is already a large amount of educational data available on the Web via proprietary and/or competing schemas and interface mechanisms, the main challenge is to start adopting LD principles and vocabularies while leveraging existing educational data available on the Web via non-LD compliant means. However, the service and data integration process is still far from seamless as different learning repositories are isolated from each other, basically existing as data silos [Dietze, 08; de Santiago, 10]. Note that in the following we use the term *metadata* in cases where we explicitly refer to resource metadata only, while otherwise, we use the more general term *data*.

Thus, four major challenges need to be addressed [Dietze, 12]:

a) Integrating distributed data from heterogeneous educational repositories: educational data and content is usually exposed by heterogeneous services/APIs such as OAI-PMH, SPARQL endpoints or SQI; therefore, interoperability is limited and Web-scale sharing of resources is not widely supported yet [Prakash, 09]. Moreover, one should also take into account the dynamic nature of available services/APIs, which may frequently be subject to additions, modifications and even removal of both content and services.

b) Metadata mediation and transformation: educational resources and the services exposing those resources are usually described by using distinct, often XML-based schemas, heterogeneous taxonomies or even unstructured text; therefore, schema and metadata transformation (e.g., into RDF) and mappings are required.

[1] http://www.w3.org/RDF/

c) Dealing with unstructured metadata: existing educational resource metadata is usually provided based on informal and poorly structured data. Free text is still widely used for describing educational resources while use of controlled vocabularies is limited and fragmented. Therefore, to allow machine-processing and Web-scale interoperability, educational metadata needs to be enriched, that is transformed into structured and formal descriptions by linking it to widely established vocabularies and datasets on the Web.

d) Shifting focus towards educational context: most currently available metadata schemes for describing educational resources focus on simplified technical and/or structural characteristics of educational content (e.g. authors, formatting, sequence of activities). A shift towards capturing the nature of the learning activities and the educational context itself is advocated [Jonassen 06].

In this paper we introduce a general approach to tackle the above challenges and demonstrate a specific implementation that aims to achieve both data and service interoperability for content sharing in the domain of health sciences education. This work has been conducted as part of the mEducator project (2009-2012)[2], an EU funded best practice network (under the eContentPlus2008 programme, Contract Nr: ECP 2008 EDU 418006) with the aim to implement and critically evaluate existing standards and reference models in the field of e-learning in order to enable specialized state-of-the-art medical educational content to be discovered, retrieved, shared and re-used across European higher academic institutions.

Our approach consists of two main steps, *(1) semantic integration*, which involves the integration of educational resources and data into a coherent data graph and *(2) social integration*, which involves the incorporation of the educational data graph into a heterogeneous social network, composed of educational resources and humans. The paper discusses related work in Section 2, while Section 3 introduces the mEducator project and illustrates the overall approach to socio-semantic integration. Section 4 details our novel approach to semantic integration of educational data, while Section 5 focuses on the social integration into an actor-resource network as part of a proof-of-concept prototype application which makes use of the proposed data and services integration approach. Section 6 provides an evaluation of the introduced approach that is discussed in Section 7.

2 Related Work

Technology enhanced learning research so far has largely focused on building and publishing comprehensive repositories of educational resources. ARIADNE (http://www.ariadne-eu.org/) was among the first consortia that significantly contributed to the sharing of training material. Several initiatives have been launched to create university-level formal education platforms providing reusable learning objects for all disciplines, while efforts are being increasingly aligned as in GLOBE (http://globe.edna.edu.au/globe/go), an international federation of a number of

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

learning object repositories, including ARIADNE, MERLOT (http://www.merlot.org/) and others.

Such repositories can only be effective if appropriate descriptions of their content are available in a standardised way. Thus, a number of official and de-facto standards and reference models have been developed to describe and manage educational content and educational content services and APIs [Devedzic, 07; Konstantinidis, 09]. Web-scale integration of educational repositories at the data and service level faces a heterogeneous landscape of Web APIs of rather isolated or only partially federated repositories. Services operating on educational repositories are rather dynamic, in that APIs appear and are removed from the Web frequently and might change behavior and interfaces according to new requirements. Therefore, it is crucial to aim at shielding the underlying heterogeneity to minimise disturbance of upper layers (e.g. educational applications). Thus, adopting service representations based on standard service vocabularies such as SAWSDL [Shet, 08] and WSMO-Lite [Vitvar, 08] is an important requirement to allow service providers and consumers to interact.

In addition, current metadata stores largely use XML and relational databases and often consist of poorly structured text lacking formal semantics, leading to ambiguous descriptions which are hard to interpret and process at machine-level. Several efforts have been made to improve interoperability by exploiting semantic technologies. For instance, early work includes efforts to provide an IEEE LOM-RDF binding [Nilsson, 03]. More recently, a Simple Query Interface (SQI) has been designed to query different learning repositories using a common query language [Simon, 05]. However, query format and result format have to be agreed among different repository providers. Recently, a peer-to-peer architecture (LOP2P) was proposed for sharing educational resources among different learning institutions [de Santiago, 10]. LOP2P aims at creating course material by using shared educational resource repositories based on a particular LOP2P plugin. A similar peer-to-peer architecture has also been proposed in the EduLearn project [Prakash, 09].

These approaches however, instead of accepting the heterogeneous landscape of the Web, impose either a common schema or interface approach on the underlying stores. Also, metadata mediation is generally based on syntactic matching, which does not adequately address ambiguities. This can be particularly important as shared educational content is rarely used in exactly the same context as it was originally created for. Semantic Web as well as Web service technologies have also been used to enable adaptation to different learning contexts by introducing a matching mechanism to map between a specific context and available learning data [Schmidt, 04]. Another similar approach, based on single shared ontology, presents scalability issues [Dietze, 08]. These issues apply as well to the idea of 'Smart Spaces' for learning [Simon, 04]. A dedicated personalization Web service proposed by [Baldoni, 06], makes use of semantic learning object descriptions to identify and provide appropriate learning content. However this work does not address integration of disparate distributed learning services or the allocation of such services at runtime. An attempt towards mediation between different educational content services is based on a so-called 'connector service' [Henze, 05; 06].

Recently, educational institutions started to expose their data based on Linked Data principles [Bizer, 09] using W3C standards such as RDF and SPARQL and emphasising the Web-based interlinking of data. This includes, for instance, The

Open University, UK (http://data.open.ac.uk), the National Research Council, CNR, Italy (http://data.cnr.it) or Southampton University, UK (http://data.southampton.ac.uk). Although this is a crucial step towards well-interlinked educational Web data, it is important to note that these efforts mainly focus on exposing data of individual institutions while interlinking with third party data is not yet within their primary scope.

In summary, current state of the art exhibits considerable work towards exposing educational content in a standardized form, mainly via conventional educational content standards and, recently, via the generic approach of Linked Data. Research is currently focusing on integrating disparate (somehow standardized) educational content repositories, although a general approach has not been demonstrated. This paper addresses this goal and proposes an approach to integrate educational content as provided by disparate repository services and expose it as Linked Data processed to support learning, in particular self directed learning. A specific implementation of this approach is also demonstrated and evaluated.

3 Overview and Approach

In this section, we describe the goals, outcomes and impact of the EC-funded best practice network (BPN) mEducator2. The mEducator consortium consisted of a highly interdisciplinary team of 15 international partners from across Europe with expertise in fields such as pedagogy and the social sciences, computer science and in the health sciences. While the general outcomes of mEducator cover a range of results, we focus here in particular on the ones generated through the following socio-semantic integration steps (see also Figure 1):

- *Semantic Integration:* Novel techniques for educational data integration (Section 4), including approaches towards repository integration, data processing and enrichment and clustering, towards improved accessibility from a learning perspective. These include tangible results such as an *infrastructure and set of APIs* (Section 4.1), a *Linked Data-compliant educational resource schema* (Section 4.2), and the *first public educational resources dataset* according to Linked Data principles (Section 4.2).
- *Social Integration:* a social application and approach that organises educational resources and actors into a socio-semantic network while exploiting the aforementioned data (Section 5).
- *Evaluation:* of the above approach for its socio-technical suitability to support learning processes (Section 6).

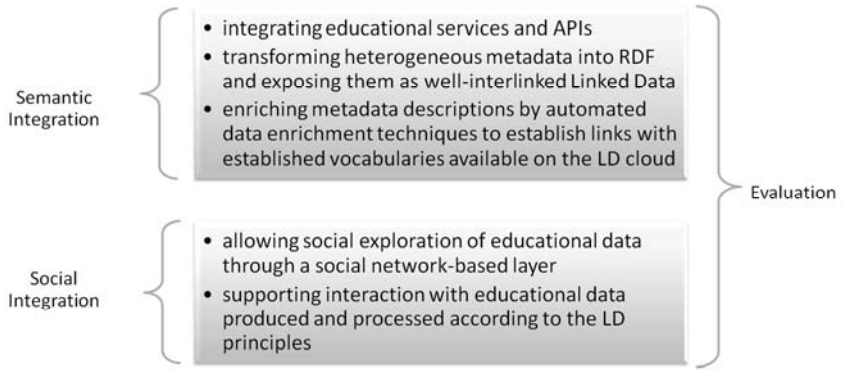


Figure 1: *Socio-semantic integration approach & overview*

To realise these steps, we have introduced the following set of technical principles and processing steps (see [Dietze, 11], [Yu, 11], and [Dietze, 12] for a more elaborate overview), which guided our developments and are briefly summarised below:

- Linked Data-principles are applied to model and expose metadata of both educational resources and educational services and APIs. In this way, not only resources are interlinked but also services and resources are exposed in a standardized and accessible way.
- Existing heterogeneous and distributed learning repositories, i.e. their Web interfaces (services) are integrated on the fly by reasoning and processing of Linked Data-based service semantics.
- Metadata retrieved from heterogeneous Web repositories, for instance IEEE LOM resource metadata or other standardized or proprietary metadata, is automatically lifted into RDF and exposed as Linked Data accessible via de-referencable URIs.
- Automated enrichment and clustering mechanisms are exploited in order to interlink retrieved data with other existing datasets as part of the LD cloud.
- Interlinked educational content items are organized and presented as social entities in a heterogeneous symmetric social network of people and content items.

Main outcomes of our work include (i) a framework (see [Dietze, 12] and [Yu, 11]) which combines a set of (ii) service/API integration methods based on SmartLink [Dietze, 11], (iii) data enrichment and interlinking methods as described in Section 4, (iv) a social environment which facilitates the embedding of educational data into an actor-resource-graph and (v) the resulting datasets as described in Sections 4-5 and evaluated in Section 6. While (i) and (ii) support the integration of educational services and APIs facilitating repository-level integration [Yu, 11], these address the challenge a) described in Section 1. The main aim is to resolve heterogeneities between individual API standards (e.g. SOAP-based services vs. REST-ful approaches) and distinct response message formats and structures (such as JSON,

XML or RDF-based ones). In order to enable integration of such heterogeneous APIs, Linked Data principles have been exploited to annotate individual APIs in terms of their interfaces, capabilities and non-functional properties (see Section 4 and [Yu, 11] for further details). All metadata of educational content retrieved from these services are transformed from their native (standardized or proprietary) formats into RDF, using a Linked Data-compliant educational resource schema. The use of this schema provides the basis for addressing the challenges b) and c) through the integration of educational resources at the (meta)data-level. Transformation of heterogeneous metadata into RDF is indeed a substantial step towards integration, however, mere transformation does not improve metadata (and data) quality. To this end, it is even more challenging to semantically integrate the lifted data by providing a common descriptive layer and disambiguate and semantically enrich data as described in challenge c). This is achieved through automated data enrichment techniques (iii) which establish links with established vocabularies (see Section 5) where an additional processing level applies machine-learning based clustering to classify the educational resources metadata into clusters that enhance the user's capability to explore the resources and afford the creation of link across different educational resources' repositories. This approach jointly with the social integration of resources into an actor-resource-graph (iv), allow educators and learners to interact with the improved educational data and resources, supporting the social exploration of and interaction with educational context, thus addressing challenge d).

4 Semantic Integration: Educational Data Integration and Interlinking

In this section, we describe the semantic integration of data, consisting of a two-fold approach: (1) integration of educational services (and, consequently, educational repositories) and (2) integration of educational data.

4.1 Educational services integration

Our current implementation builds on existing research and uses rather lightweight service annotation schemas to apply Linked Data principles to the services domain. Based on RDF descriptions of core elements of services and APIs, these are discovered and executed according to a given set of service consumer constraints. We exploit two well-integrated technologies which follow Linked Data-principles for services and API annotation and integration: iServe [Pedrinaci, 10] and SmartLink [Dietze, 11] are two public LD-based environments dealing with two kinds of service annotations separately, namely, functional and non-functional service annotations stored in dedicated RDF stores. SmartLink has been partially developed and in particular, evaluated, as part of the mEducator project. Both iServe and SmartLink adopt LD principles to expose services and APIs, expressed in terms of simple conceptual service models, based on established schemas such as SAWSDL [Sheth, 08], FOAF or CommonTag (http://commontag.org/home), to enable consumption by both humans and machines. In addition, SmartLink provides a Web-based interface which allows annotation of services and browsing of existing descriptions within the SmartLink and iServe repositories via a unified user interface.

Initially, educational repository service providers annotate services from scratch (i.e. without any pre-existing services documentation such as WSDL or HTML files) and expose RDF descriptions. Published services[3] can then be discovered via a set of RESTful APIs. Identified services are invoked and heterogeneous service responses are lifted into a coherent RDF schema. This is achieved on the basis of *lifting* and *lowering* templates, which are part of each, service description ([Dietze, 11] and [Pedrinaci, 10]) and are usually created by the service provider or annotator. These templates specify how, based on a SmartLink service invocation message, the particular invocation input parameters are lowered into the representation required by the particular service and how response messages are lifted into RDF statements according to the specified lifting schema.

For this particular implementation in the domain of health sciences education, the unified mEducator RDF schema [Mitsopoulou, 11] was developed, to address specifically medical educational resources. A number of predicates are part of other popular namespaces, e.g. the Dublin Core (ISO 15836:2009) Elements and Terms namespaces (http://dublincore.org/documents/dces/) for a generic resource description, FOAF (http://www.foaf-project.org/) for describing people, SIOC (http://sioc-project.org/) for integrating online communities and SKOS (http://www.w3.org/2004/02/skos/) for knowledge organization, while a number of novel domain-specific properties are also included.

Our current approach uses a number of existing vocabularies, for which in many cases mappings already are or could be provided. SmartLink/iServe use a range of established schemas, and import mechanisms, for instance to import OWL-S-based services are also provided. While it is an inherent aspect of the LD-approach that different schemas and datasets are gradually mapped and interlinked, we particularly envisage also alignment with other schemas which might be deployed be individual service providers. The strength of the LD approach is its inherent ability to reason and consider such links (for instance, owl:sameAs links or inheritance of classes and properties across schemas).

Based on service input and output annotations and corresponding *lifting/lowering* schema descriptions within SmartLink/iServe, an RDF service invocation message is generated. This is done by dynamically matching service execution input parameters (specified by service consumers) to semantic service input descriptions. For instance, a *keyword* such as *thrombolysis*, which is used in an educational resource search, will be mapped to the RDF description of the service input during the invocation process. The RDF input message will pass a *lowering* process to generate the actual input format of the service. A service is then invoked with the lowered input.

In order to provide service response messages compliant with the mEducator RDF schema, native output from service invocations (e.g., XML, JSON) are transformed via the *lifting* step to RDF compliant with the mEducator RDF schema. The lifting step uses XSPARQL (http://www.w3.org/Submission/2009/01/) to lift XML vocabularies to RDF triples. A dedicated RESTful API allows third party applications to interact with the RDF annotations of educational services and APIs, for instance, to discover and execute services (the API can be accessed at

[3] Services are continuously updated. The current list of services can be retrieved dynamically from SmartLink via the following request:
http://smartlink.open.ac.uk/servicerestapi/restapi/searchservices

http://smartlink.open.ac.uk/servicerestapi/restapi/). Educational metadata as retrieved in the services integration step is stored in a dedicated RDF store containing the *mEducator – Linked Educational Resources* and implemented based on Sesame/OWLIM (http://www.ontotext.com/owlim/). A dedicated REST API is offered and each resource entity, described using the mEducator resource schema, owns a unique, dereferencable URI.

4.2 Educational data enrichment and interlinking

Educational metadata retrieved and exposed according to the previous steps is most often poorly structured as it is based on either unstructured text or often less well-defined terminologies. Thus, such metadata needs to be enriched so as to expand it with additional information and ensure disambiguation, as well as to allow correlation with similar or related information and resources. In particular, the Linked Data cloud already offers large amounts of datasets, ranging from general-purpose ones to domain-specific educational datasets. To this end, we have developed enrichment mechanisms that automatically enrich poorly structured descriptions with links to related terms in well-established vocabularies. 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 the general purpose DBpedia Spotlight entity extraction and semantic annotation tool (http://spotlight.dbpedia.org); and the medical domain specific BioPortal [Noy, 09]. The latter is an open repository of biomedical ontologies that allows access to a vast number of established taxonomies and vocabularies, such as SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms), ICD9/10 (International Statistical Classification Diseases and Related Health Problems), Body System (body system terms used in ICD11), MeSH (Medical Subject Headings), NCI (Meta)Thesaurus, Galen (the high level ontology for the medical domain), HL7 (the Normative RIM model v2), Biomedical Resource Ontology (BRO, a controlled terminology of resources to improve sensitivity and specificity of Web searches).

Enrichment is implemented currently in two different ways: (a) as an automated mechanism whenever new data is 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, the second approach makes exclusive use of the BioPortal API.

Unstructured free text (as part of resource metadata), for instance the keyword “thrombolysis”, is enriched with unique URIs of structured LD entities - such as *http://dbpedia.org/resource/Thrombolysis* 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. Such enrichment allows further reasoning on related concepts, enables users to query for resources by using well-defined concepts and terms as opposed to ambiguous free text, and enables disambiguation.

For instance, to refer to another example from our established dataset, a particular resource contains a learning objective description stating “1. Better understanding of the role of HPV in cervical carcinogenesis 2. Better understanding of the natural history of HPV infection 3. Better understanding of the benefits and limitations of

HPV testing". The automated enrichment mechanism correctly associates the DBpedia enrichment *http://dbpedia.org/resource/Human_papillomavirus* by exploiting the context of the term "HPV" – which is also used, for instance, as abbreviation of Human Powered Vehicle – providing proper classification and disambiguation of the resource description. Subject identifiers (URIs) are automatically generated based on a mEducator-specific namespace and the URI of the external resource (e.g. the corresponding DBpedia reference). This allows modeling unique instances within our dataset which do not redefine DBpedia resources but merely are linked to them via *rdfs:isDefinedBy* relations and carry additional meta-information about the enrichment.

Although a tremendous amount of effort has been devoted by the medical informatics community, a universal standard medical terminology and/or ontology has yet to be achieved [Lenz, 07]; thus current state of the art presents a wealth of numerous diverse and often overlapping vocabularies, terminologies and ontologies (e.g. more than 340 available only in NCBO Bioportal). Research efforts have focused on producing mappings between different terminologies, probably the most notable contribution is the UMLS mapping between more than 100 of the most commonly used controlled vocabularies in healthcare [Bodenreider, 04]. In order to circumvent this problem, our approach employs a semi-automated enrichment, which only recommends suitable entities from BioPortal, while the user has full control to filter and select only entities retrieved from selected and desired vocabularies (for instance SNOMED-CT or MeSH), which express the intended semantics. In addition, the LD approach particularly foresees the introduction of links between disparate datasets and vocabularies, what can help in qualifying the relationship between related enrichments. While it is out of scope of our work to interlink all existing and used vocabularies, the inherent facilities of LD facilitate the tighter interlinking of domain-specific vocabularies.

Based on the common structured concept references as generated via enrichment, resources from different repositories can also be implicitly clustered. Starting from the relationship between mEducator resources and DBpedia/BioPortal concepts, we created a weighted graph (Figure 2) in which mEducator resources are represented as blue square nodes and enrichments as green circle nodes. The weight of each edge identify the strength of the relationships between nodes, and these features have been exploited by the clustering algorithm to aggregate resources. Considering the resources and their links with structured concepts from, for instance DBpedia, distinct resources often share at least one common concept reference, what facilitates correlation of resources. Therefore, we implicitly provided a means of clustering resources originating from distinct repositories, with the aim of producing new connections between similar resources that were not explicitly connected before. In the figure, connections between resources sharing the same enrichments have been discovered: for instance the DBpedia enrichment for "*chest pain*" create new links between resources not explicitly connected, providing the mean to aggregate similar resources in clusters. In order to create more coherent and cohesive clusters of nodes we analyzed the graph using different thresholds for number of common concepts (see Section 6.2).

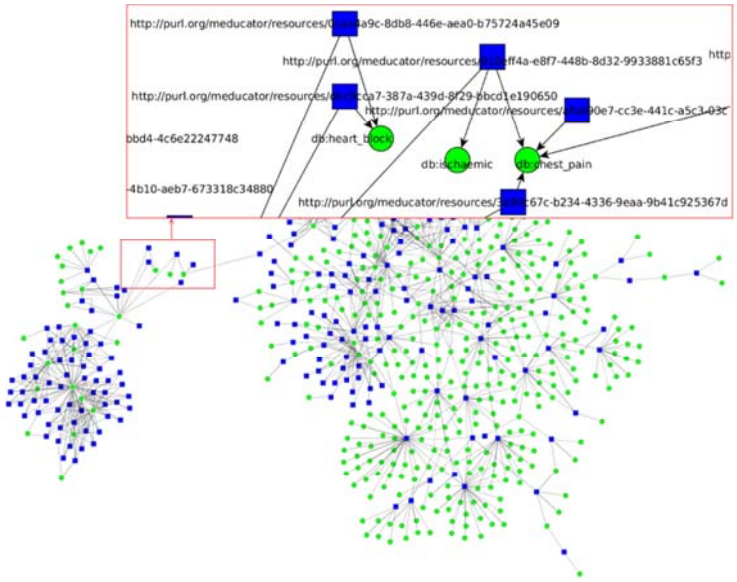


Figure 2: Relationships between educational resources from distinct repositories based on shared DBpedia references(an edge denotes that there is at least one link between the two nodes, i.e., the resource and its enrichment)

4.3 Clustering and exploratory search based on linguistic similarities

It has been recently recognized that users often engage in a search process, known as exploratory search [Marchionini, 06]. Exploratory search is characterised as being open-ended. It involves learning and refocusing the information need while the user makes sense of the information retrieved after sequences of tentative queries that are progressively adjusted. Exploratory search is starting to be considered in the Semantic Web community, for example by addressing issues of how to design exploratory browsing tools [Kobilarov, 08].

In our work we employ an exploratory search approach proposed by [Giordano, 09] which is based on the use of a multiple unsupervised clustering technique that uses as a basis users' selected subsets of the mEducator metadata fields. This method also derives "second order" relationships among items from the analysis of the defining features of each cluster: when commonalities across clusters are found above a given threshold, association rules among the defining features can be drawn; these may be used to further expand the recall of a user query's results, but based on new, different "facets". Whereas the terminology enrichment process discussed in section 4.2 tends to increase the precision of retrieval, and thus is key in supporting focused search, the use of clustering based on some notion of similarity plays an orthogonal role: it supports exploratory search (a) by increasing the recall of retrieval, and (b) by suggesting possibly relevant items, as found in the clusters that are proposed for exploration in association with the regular query's results.

The clustering functionalities have been implemented to allow the interlinking of resources originating from different repositories. In particular, the clustering process is based on three main phases: content indexing, creation of similarity matrices and clustering itself. The first step parses the RDF fields of the selected subsets of the unifying metadata schema, including those that contain descriptive free text, extracts the keywords by applying an inverse term frequency approach [Robertson, 04], stems the terms and starting from the keywords applies the Random Indexing method [Sahlgren, 05] with a configurable sliding window. The output of this step is a multidimensional matrix, namely, Doc-Term (DT) that contains the frequency of each keyword in a given document (metadata instance), together with the contexts in which each keyword occurs. Each context is a vector of neighbouring terms, with size same as the sliding window. Once the indexing phase is completed, in the second phase, a similarity matrix S is created, by assigning a weighted similarity score among the resources. The S matrix then is used to classify the resources, by applying a clustering algorithm; our current implementation supports both K-means and clustering based on Kohonen self-organizing maps [Kohonen, 95].

We are currently investigating appropriate metrics to analyze the provenance of the items in the clusters to derive semantically qualified associations at the repository level, based, for example, on subject coverage similarity, or on content type similarity. An important challenge is to make the method scalable with the number of educational resources annotated and integrated in the RDF store, especially with respect to the computational aspects of indexing and re-clustering periodically necessary to take into account changes in the constellation of available resources. In this respect, a viable solution is the method presented in [Faro, 11] that makes use of parallel, distributed computation. This approach may eventually support faster and more sophisticated automatic repositories interlinking, and enable real-time re-clustering of the items based on end-user specific requests, which is a current frontier for exploratory search.

5 Social Integration - Educational Resources as Social Entities in Metamorphosis+

The data and services integration APIs and datasets presented in the previous sections are fully integrated in the MetaMorphosis+ (http://metamorphosis.med.duth.gr/) environment, which merges the paradigms of semantic and social web to produce an environment for sharing linked educational resources.

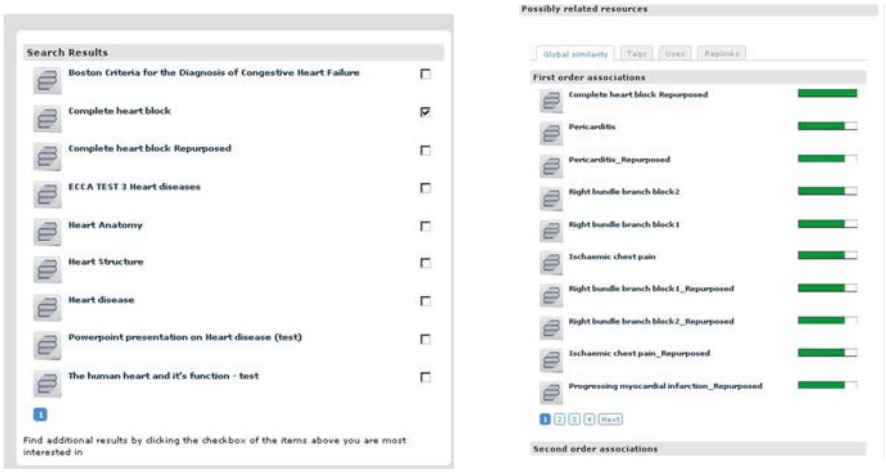
The social Web, or Web 2.0 [O'Reilly, 07] has become an important trend during the last few years. Among the prominent social web tools, social networking websites focus on creating online communities of individuals who publish their content and activities while exploring others content and activities, thus creating virtual online social groups and associations. A conventional social network approach concentrates on the network of humans, presumably based on some common social or professional interest, but it doesn't explain what connects those particular people together and what connects those and not others [Knorr-Cetina, 97]. Recently, the term 'object-centered sociality' was introduced [Engeström, 05] to describe the fact that strong social relationships are built mainly when individuals are grouped together around a

shared object that mediates the ties between them. This can be achieved by organizing the network around the content people create together, comment on, link to, annotate etc. [Breslin, 07]. This new approach to sociality has drawn attention, and current state-of-the-art research in the area involves various ways to exploit object-oriented sociality to the benefit of the community. Some indicative examples from social networking in educational settings include Edmodo (http://www.edmodo.com/) a social network for teachers and students who can interact in private virtual classrooms to share educational content and activities; Brainify (http://www.brainify.com/) an academic social bookmarking and networking site for university students and professors, mainly built around commonly collected and shared website links; and wePapers (http://www.wepapers.com/) a network organized around papers, lecture notes and other documents students and teachers share.

In both cases of human-centered and object-centered networks, the focus is on the creation of associations based on human action, agency and perception, thus placing always emphasis on human agency. In MetaMorphosis+ we follow a different novel view for truly heterogeneous social networks where humans and nonhuman entities (i.e. educational resources in this case) are integrated into the same conceptual framework and assigned equal amounts of agency [Kaldoudi, 11a]. In implementing such a network, major challenges include a unified treatment and representation of all types of possible actors as well as the development of a social behavior for various nonhuman actors, and subsequently their own associations and networks. Both challenges can be addressed by concepts and technologies of the Semantic Web as described in previous sections.

In general, the social aspect of non-human actors in MetaMorphosis+ is created in a variety of ways, including [Kaldoudi, 11b]: (a) the obvious connections via common tags that are used in their profile description; (b) connections based on collective usage and other related interaction of human users, i.e. what human users do with the nonhuman entities; (c) social connections based on some type of inheritance, i.e. non-human entities that are generated or are the product of other resources, in the sense of the genealogy tree; and (d) semantic connections and similarities that can be built based on the wealth of information available in the Linked Data cloud.

The first aspect of educational resource sociality is built based on obvious connections and shared annotations between different educational resources, as these are created via the automatic and/or semi-automatic metadata enrichment mechanisms described earlier. The user entered keyword is used to retrieve all relevant standardized concepts from the medical ontologies/vocabularies available via NCBO BioPortal, using either the entire BioPortal collection or only specific vocabularies specified by the user. The retrieved concepts can then be used to enrich specific metadata fields as indicated by the user. Our previously described clustering mechanisms are exploited to enable exploratory search and navigation through educational resources. For example, the screen shot in Figure 3 (a) shows a set of resources retrieved via simple keyword based search (e.g. for the keyword ‘heart’) using the RDF store APIs. The screen shot in figure 3 (b) shows the associations created for one specific resource (in this case, the second result in Figure 3.(a) with other resources based on the clustering mechanism described above. A visual indication for the degree of association is also included.



(a) *(b)*

Figure 3: Basic search mechanism in MetaMorphosis+ and exploratory search results enabled via clustering of resources based on the introduced mechanisms

Finally, an important aspect of social connections relates to resource family trees based on repurposing history and inheritance. The term 'repurposing' refers to changing a learning resource initially created and used for a specific educational purpose in a specific educational context in order to fit a different new educational purpose in the same or different educational context. Although not formally addressed as such, educational content repurposing is what any educator is routinely engaged in when preparing a new educational experience, including preparing the educational content itself. Customarily, when an educator sets the context and goals of a new educational experience, he/she will overview existing content and/or search for new relative content and then repurpose and re-organize content to fit the purpose of the new educational experience. In MetaMorphosis+ repurposing is addressed as a means to provide a different kind of sociality for the educational resources. Thus repurposing history and inheritance are used as basic social relationship among educational resources in order to cluster resources into families. Each repurposed resource declares its parent(s) resource(s). Following iteratively the 'parents' in a chain of repurposing ancestors, the entire 'family' tree of the particular resource can be compiled.

Considering all possible social connections for any particular educational resource, its individual actor-network can then be built, highlighting its authors, other users that somehow interact with the resource, the resource repurposing family and any other connection via semantic linking. An example of such a rich individual actor-network is shown in Figure 4 for the resource "Climate Health Impact Game", highlighting the resource repurposing family as well as any other social connections (up to the 3rd degree) which include connections with humans and other resources alike.

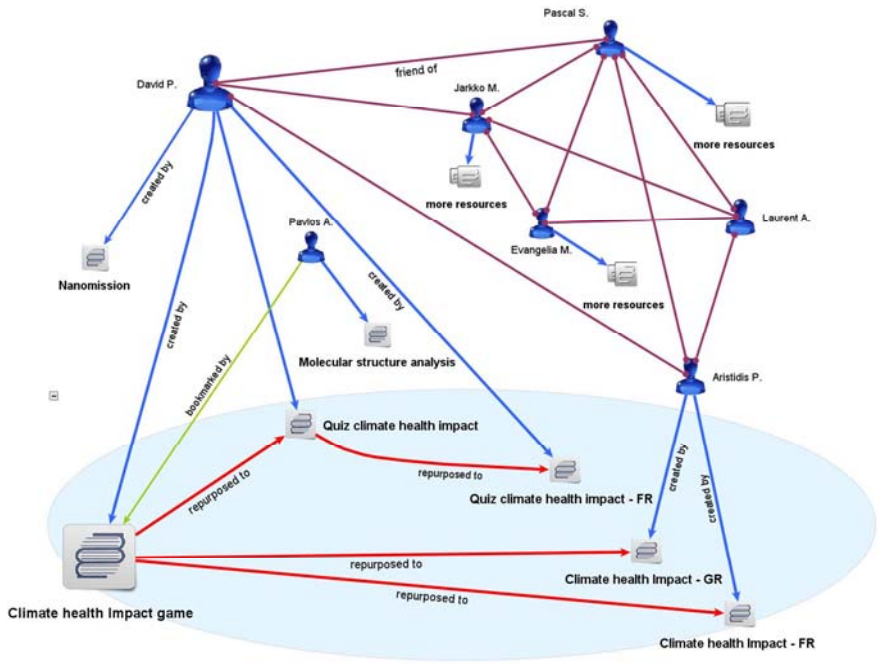


Figure 4: Basic Actor-network graph for a specific educational resource ('Climate Health Impact Game') that shows connections of variable types (up to the 3rd degree)

MetaMorphosis+ is implemented using the *Elgg* open source social engine (http://www.elgg.org/) appropriately modified to allow for the formation of a heterogeneous human/resources network as well as the exploitation of the underlying semantic framework as presented in previous sections. MetaMorphosis+ is used by a total of 560 registered users (as of September 25, 2011). Almost half of the resources (43%) are in English, while there is a representation of more than 15 other European languages. The majority of resources (80%) have metadata described in English language, while the rest have metadata in other languages. The resources included in the environment are distributed among the various educational levels, 25% intended for undergraduate medical education, 17% intended for postgraduate/resident studies and 19% for continuing life-long education, while 19% are intended for educating the public.

6 Evaluation: data quality and clustering added value assessment

In this section, we provide an evaluation of the data quality of the obtained resource metadata within the *mEducator Linked Educational Resources* dataset, and an assessment of the added value provided by the clustering functionalities.

Several approaches could be adopted to evaluate the data produced through our suggested approach for socio-semantic integration of educational resources. Strict focus on support for the learning processes as envisaged in the mEducator project, i.e., self-directed learning of students in the health sciences and repurposing of learning resources by teachers and tutors, would require extensive user studies of acceptance of certain features together with an assessment of their utility, as and a prolonged observation to track actual usage and learning goals. Whereas this approach can definitely provide insights into the assumptions about the learning processes that underly the design of educational content sharing systems, as is MetaMorphosis+, the results of this type of evaluation could be potentially biased by the specifics of the design (e.g., user interface design choices, overall usability).

In the context of this work we are interested in evaluating more objectively the aspects of our methodology that we posit to: 1) improve overall resource interoperability (both at the syntactic and semantic level), 2) effectively enrich a given resource, by disambiguating its terminology and by providing links to meaningfully related resources, and 3) widen the scope of the related resources retrievable by the users.

Therefore, the central aspects we are investigating are:

(A1) *interoperability gains*, operationalised by the number of resources that were lifted into a unified knowledge graph, and by the number of new links established via shared enrichments (Section 6.2).

(A2) *accuracy of introduced links*, as measured by the precision of enrichment and its semantic disambiguation (Section 6.3).

(A3) *performance of machine learning-based clustering*, in terms of its precision and recall (Section 6.4).

A1 is assessed with a quantitative analysis of generated data, while A2 is addressed based on a qualitative, semantic assessment of the enriched data, and of the enrichments in particular, together with an evaluation of the clusters generated by taking into account shared enrichments. In addition, we provide a performance evaluation of machine learning clusters to address A3, together with an assessment of the potential for using the clusters as sources for interlinking resources and support exploratory search. A1 relates to challenges a) and b) - identified in Section 1, A2 addresses challenge c), whereas A3 relates to challenges a) - as it also contributes to generate links across resources for distributed, heterogeneous resources, and, partly, to challenge d). In fact, the challenge to take into account the actual educational context of a learning resource is catered for in MetaMorphosis+ by the social network functionalities, and by the object-actor repurposing network. However, machine learning clustering, as it will be shown in section 6.4, handles some implicit aspects in the linguistic description of resources that are potentially useful for an indirect characterization of the educational context of usage for the resource.

An evaluation of the performance - from a strictly technical perspective - and scalability of our approach is presented in [Hendrix, 12].

6.1 Dataset for evaluation

The evaluation is based on the educational resources dataset as of September 2012 which exhibits 375 resources with the following overall characteristics:

- educational level of resources: 25% undergraduate, 17% postgraduate, 19% continuing education, 19% patient-public education, 4% other;
- language of educational resources: more than 15 languages, 43% English, 18% Bulgarian, 7% French, 6% Greek, 2% Romanian, 2% Italian, 2% Finish;
- resource types: conventional (text, presentation, notes) 49%, multimedia 16%, case/problem based material 12%, web 2.0 4%, serious games 2%;used by educators and learners from more than 20 different countries, from various related fields: 90% medical, 5% medical informatics, 4% biology, 1% medical physics.

6.2 Quantitative data assessment

Note, that no automatic mass imports of existing metadata resources were conducted, but all data was (a) retrieved on-the-fly by queries sent to distributed stores via the approach described in Section Educational Services Integration and then (b) enriched automatically based on the approach described in Section Linked Data-based Enrichment & Clustering. Therefore, baseline metadata for each resource was iteratively retrieved from existing repositories and automatically transformed into the mEducator resources schema and enriched with LOD references within the RDF store.

At the date of evaluation, the *mEducator Linked Educational Resources* RDF store contains in total 23876 triples, of which 10206 directly refer to a total N=375 distinct educational resources. The average number of triples per learning resource is 27, ranging from a minimum of 6 to a maximum of 68.

In addition, it has been shown 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 1 provides an overview of property usage frequency across educational resources. Properties are listed in order of decreasing frequency. A more detailed description of individual properties is given in [Mitsopoulou, 11].

Property usage frequency across educational resources					
Property	**%**	**Property**	**%**	**Property**	**%**
mdc:creator	100	mdc:hasEnrichmentContext	79.2	mdc:mediaType	37.6
mdc:description	100	mdc:rights	77.9	mdc:isRepurposedFrom	24.6
mdc:metadataCreated	100	mdc:discipline	76.5	mdc:repurposingContext	22.13
mdc:metadataCreator	100	mdc:technicalDescription	68.3	mdc:repurposingDescription	19.7
mdc:title	100	mdc:educationalContext	66.7	mdc:quality	1.9
mdc:created	95.7	mdc:educationalLevel	64	mdc:isAccompaniedBy	0
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		
Range [100-80%]		**Range [80-40%]**		**Range [40-0%]**	

Table 1: Resource schema property usage frequency across educational resource

Data collected in the RDF store makes use of a set of external LOD vocabularies/datasets (see above). In detail, 545 distinct terms have been associated based on the automatic and semi-automatic methodologies described in the previous sections. Table 2 shows the distribution of the used terms by external sources, and the correspondent percentage.

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.

For 297 resources (i.e. 79.2% of the total number of resources) enrichment data generated through linking to DBpedia is available. The number of enrichments in the data store is 1352, involving 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 our automated enrichment based on DBpedia Spotlight, while the semi-automated approach via the BioPortal API provided a higher diversity – data from different vocabularies such as MESH and SNOMED are used – but only a very limited amount of overall enrichments, since it requires manual intervention and pre-selection of suggested terms. Most highly used

properties for enrichment are *mdc:description* (54.2 %), *mdc:title* (24.2 %), *mdc:educationalOutcomes* (21.6 %).

The automated enrichment procedure has linked 303 of the 375 resources to 502 different DBPedia concepts. The following table represents the number of enrichments in relation to the number of resources. The 20 most used enrichment concepts are listed in the table below. The majority of resources have between 1 and 5 enrichments, but there are also resources that were linked to over 20 DBpedia concepts. The most enriched resource is linked to 42 different DBpedia concepts. The enrichment procedure has, thus, created new connections between resources sharing the enrichments. The resulting graph is dense and interconnected, providing a positive measure of the semantic interoperability gain supported by our procedure.

DBpedia concept (http://dbpedia.org/resource/....)	Linked by number of resources
Cervical_cancer	59
Screening	31
Cervical	29
Hpv	29
Oxygenation	26
Childhood	22
differential_diagnosis	19
Knowledge	18
Learning	17
decision_making	16
Training	15
Lecture	15
Risk	15
hpv_infection	15
Fear	15
pap_smear	15
Abnormal	14
Ventilation	14
Ecg	14

Table 3: Frequency of DBpedia references

6.3 Qualitative evaluation of enrichments

To evaluate the quality of the automated enrichment procedure results, a set of 200 enrichments from the overall set of 1352 has been selected randomly, and assessed manually by experts. Based on their assessment, the experts assigned one of 3 possible ratings to each enrichment to indicate its grade of meaningfulness:

(A) The description of the external enrichment concept describes and expands the semantic meaning of the enriched source text.
(B) The enrichment concept shows a significant relationship with the enriched text but does not provide a direct mapping. This category also was assigned to

enrichments where the correctness could not be determined with certainty due to ambiguity of the source concept.

(C) The enrichment has no meaningful relationship with the source concept.

The evaluation has led the following results: 92% of enrichments belong to the category (A), while the 8% of enrichments is equally distributed between the categories (B) and (C).

Assigning a value of 1 for the category A, 0.5 for the category B and 0 for the category C the overall average score for the selected enrichments is: 0.94. While this value provides a positive evaluation result, we also identified some minor amount of noise contributed by the current enrichment approaches. We are currently investigating ways to reduce the amount of less accurate enrichments by, for instance, expanding the amount of considered context used by the enrichment queries and by replacing DBpedia Spotlight with typed queries to DBpedia.

While the analysis did not yet assess in a structured manner the correctness of the generated clusters, it has confirmed its potential for disambiguation by associating acronyms and synonyms with enriched text. For example, description of medical learning resources showing the ambiguous acronym *HPV* have been correctly enriched with references to *http://dbpedia.org/resource/Human_papillomavirus* based on the co-occuring terms in the resource description. Similar behavior has been detected with the acronyms *TMJ* (Temporomandibular joint) and *EASA* (European Aviation Safety Agency). Concerning synonyms, descriptions containing the words: *pap smear* or *gastric cancer* have been respectively enriched with corresponding *pap_test* and *stomach_cancer* concepts within DBpedia. Such kinds of enrichments allow broader yet more precise search and clustering results, which will be subject to further investigation. In order to assess the quality of the derived enrichments, we analysed the clusters retrieved from grouping resources with equivalent enrichments (DBpedia in this case). To this end, we generated different views on the generated graph by clustering resources according to different thresholds *T*, where *T* defines the minimum amount of equivalent enrichments of all resources *R* within a particular cluster *C*. For instance, applying a threshold *T=2* led to 13 clusters in which all resources shared at least 2 enrichment concepts. In contrast, less but more coherent clusters were created with *T=4*, that is, there is an edge between two nodes only if they have more than four concepts in common.

It is important to note that the clusters with less than two resources have been removed from our analysis. The following table shows the nodes a cluster of the graph generated with T=4. Manual assessment proved a strong semantic correlation of the clustered resources, i.e. the majority of the enrichment concepts used in this cluster are related to similar diseases or disease aspects (cervical cancer, screening).

Based on these observations, it can be stated that the obtained resource enrichments from LOD vocabularies prove a useful means to correlate and cluster heterogeneous references. While lower thresholds lead to higher level of noise within individual clusters and hence, less cohesive resource clusters, current research is dealing with identifying the most suitable threshold value.

	screening	abnormal	cervical_smear	cervical_cancer	fear	pap_smear
80986a54-dc69-441e-b2b9-37538c0c893b	1	0	0	1	1	1
a13d97c1-0b4c-4553-bf87-329ad318367a	1	0	0	1	1	1
59c56028-fc24-4e35-abcc-22c078f58f51	1	0	0	1	1	1
0349caad-06cf-4557-b5fd-da09fc349b2b	1	0	0	1	1	1
25e5b2b8-13e9-4415-bb73-6332e9e6bd03	1	0	0	1	1	1
30a799f0-e443-4d25-821e-2bdf47d2aa92	1	0	0	1	1	1
02d248c2-dd1b-4edc-a7ea-ec4137737976	1	0	0	1	1	1
220007f5-0927-4715-a2c8-03d1ab7cb8f5	1	0	0	1	1	1
48d7cce9-3b42-455e-876f-89aef831bbc7	1	1	1	1	1	1
5890fd27-96a2-4e66-96e6-3983e754cd69	1	0	0	1	1	1
32c05491-76fd-4f9f-a6f5-b493ac69107a	1	0	0	1	1	1
74ffc536-87d9-4ca3-93c7-7e579b2f0ccd	1	0	0	1	1	1
aedc8357-b72d-4c34-99b5-20f1e72ac1bd	1	1	1	1	1	1
de79f455-2466-4947-a131-365de8733028	1	0	0	1	1	1
f2c81f0e-17ce-45f9-971b-f6ca3623180c	1	0	0	1	1	1

Table 4: Example cluster, enrichments/resources (T=4)

6.4 Performance evaluation of machine learning clustering

The clustering based on the machine learning approach outlined in section 4.3 is evaluated from two perspectives: potential for interlinking and support for exploratory search. The results of clustering are notoriously difficult to evaluate, because several classifications could be equally valid on the same dataset, and because clustering performance also depends on the nature of the dataset, and on the specific notion of similarity that is adopted [Grimnes, 08]; when a ground truth is available the standard measures to assess the clusters are precision P, recall R and F-Measure, which is the harmonic mean of precision and recall. We here focus on the evaluation on two reference cases: 1) clustering based on unsupervised K-means in which each resource in Metamorphosis+ is classified only in one class, and 2) a clustering method according to which a single item can belong to several classes. By the results of the standard k-means clustering, performance of our linguistic similarity metric with the specific dataset and the potential for interlinking across resources across repositories are assessed. Multiple clusters classification is used to evaluate the support for exploratory search, by measuring the increases in recall with respect to a set of queries that operationalize an exploratory search goal. The resources in the dataset were classified indipendently by two subjects: a computer scientist and a medical doctor. Based on the available metadata, they were asked to group the dataset first in 5 clusters, then in 10 clusters and in 15 clusters. A standard K-means was run at 5, 10, and 15, using the following metadata fields: Title, Description, Discipline, DisciplineSpecialty, and Keywords. The results are shown in Table 5.

Table 5 shows a good coherence of the obtained clusters, since the average F-measure ranges from 0.74 (5 clusters) to 0.55 (15 clusters). Typical F-measures obtained by k-means are in the range 0.12-0.13 for the Citeseer dataset and 0.42-0.44 for the PIMO (Personal Information Model from a sematic desktop system) dataset [Grimnes, 08]. The results show that the adopted similarity measure performs well on the rather concise descriptions of the used metadata (as opposed to full text indexing).

	Subject 1			Subject 2		
Clusters	**P**	**R**	**F**	**P**	**R**	**F**
5	0.84	0.81	0.83	0.63	0.68	0.66
10	0.84	0.84	0.84	0.45	0.56	0.50
15	0.62	0.70	0.66	0.42	0.48	0.45

Table 5: Average Precision, Recall and F-Measure for Subject1 (Computer Scientist) and Subject2 (Medical Staff)

Table 6 reports the average number of repositories contained in each cluster, and on average, how many resources from each distinct repository are contained in each cluster. The number of resources tends to stay stable across different classifications, whereas the finer grained classification of 10 or 15 clusters appears suitable to discover potential interlinks across repositories.

Clusters	Average Repositories per Cluster	Average Resources per Repository in Cluster
5	9.83±4.54	2.16±1.22
10	4.91 ± 2.64	2.19±1.21
15	3.03±1.26	1.97±0.94

Table 6: Average distinct repositories per cluster, and average resources per repository in cluster

To assess the potential for using the information contained in the cluster to interlink resources, a test was performed to highlight those associations that are stable regardless of the number of clusters chosen for the classification. The test points out two interesting cases. First, a stable association was found between Carotid Stenosis and Raynaud's Syndrome, for which there are a few recent studies that point to a correlation between the two diseases. These resources are from the same repository. This finding exemplifies the potential of the system from the discovery/hypothesis generation perspective. Although scientific discovery is not the main purpose of the system, the implied educational implication is clear, since cluster analysis can stimulate the learners in generating questions, hypothesis and in the making of conceptual connections. From a pure classification point of view, whereas at 5 clusters all the "Virtual Patients" resources are grouped together, at 15 clusters the method is able to differentiate those resources dealing with emergency cases from those dealing with pediatric cases. Also, the extracted stable associations reveal clusters that tend to group resources from the same repository (although provenance was not intervening explicitly in the classification) and clusters that provide a transversal, conceptual classification; notably, one set of stable associations features diagnostic procedures, and yet another one groups resources featuring uses of information technology in education and medicine. These clusters are the most varied

in terms of repository provenance. At the time of testing, the data set was still too limited to investigate other metrics that might support the linking at the repository level, beyond the links that can be established at the individual resources' level, as demonstrated by the stable associations analysis.

To assess exploratory search, we have simulated a search to collect resources for possible usage in a course in medical informatics, operationalized through a set of search terms that include general compound terms (e.g. medical informatics, information technology), typical acronyms (e.g., ICT, EHR), specific medical informatics concepts (e.g., standards, telemedicine, etc.). The dataset was annotated w.r.t. the relevance of the resources to the proposed queries by one of the two independent subjects. By running a multiclass clustering method taking into account all the metadata fields ("global similarity") the new relevant items retrieved by exploring the cluster and that would have not been retrieved by the set of queries based on the selected search terms were counted, to estimate the increase in recall. With respect to the baseline results computed by using the "title" property in the Metamorphosis + Advanced search form (precision 83% and the average recall 27%) the increase factor in recall with multiclass clustering was 2,4 (from 27% to 65%). This is obtained without excessive noise in the clusters (average precision is 60,9%), which means that the users can inspect them without much effort.

7 Discussion and Conclusion

Integrating existing educational Web resources becomes increasingly important as plenty of metadata and related data is published openly online with the aim of Web-scale sharing and reuse of resources. We have proposed an approach that fundamentally builds on exploiting Linked Data principles to support Web-scale interoperability between educational resources, that is, educational services as well as data. In particular, we tackle *Semantic Integration* as well as *Social Interconnecting* of educational data, resources and actors. By exposing educational resources via Linked Data principles, we leverage on the wealth of existing datasets and vocabularies and allow interlinking between educational data and resources. We have introduced a set of implemented integration approaches, resulting RDF datasets and APIs, and an application (MetaMorphosis+), which provides an open environment for (biomedical) education to aid teachers and learners in the field of biomedicine and which showcases the use of our integration mechanism for Linked Education. Exploiting our APIs for educational data and services integration enables previously unavailable features, such as Web-wide search of educational resources (via SmartLink), exploratory search based on resource clusters and more precise search results based on the use of established and semantically rich vocabularies.

The quantitative analysis described in section 6.2, has reported that the number of additional generated links (considering shared enrichments), is considerably increased, thus leading to an interoperability gain. Qualitative evaluation of the established links confirmed only insignificant amounts of noise, but further investigations are needed, due to the limited amount of data generated so far. Current work is replacing the semi-automated data import (via SmartLink) with periodical data harvesting mechanisms to significantly increase the amount of integrated data. Results of qualitative assessment of the enrichment and clustering techniques indicate

sufficient precision values to confirm the suitability of both approaches to enrich integrated OER Web data in a way which enables more exploratory navigation mechanisms for end users. The performance evaluation could confirm acceptable response times, however, it uncovered significant issues with respect to scalability.

While the presented work already tackles a number of distinct challenges such as metadata interoperability, services discovery or data mediation, pending issues uncovered during evaluation activities will be addressed as part of ongoing and future work. Most importantly, the limited amount of generated data, as well as scalability and performance issues will be addressed by replacing semi-automated data imports with a periodic data harvesting mechanism from all integrated data stores. Thus, we will establish a unified entry point to well-interlinked educational datasets on the Web based on the principles described in this paper. In addition we will: (1) investigate additional ways to enable efficient, accurate and dynamic enrichment of educational data, by resorting to techniques such as text mining, entity recognition and ontology mapping; (2) extend our framework with additional educational data stores to further evaluate our services integration approach; (3) integrate the APIs introduced in this paper with additional third party applications to further evaluate the performance and scalability of our approach. Finally, while some aspects of our proposed framework are domain-independent, the deployment of similar approaches in areas different to education, such as eScience, is yet another focus of investigation.

Acknowledgements

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

References

[Baldoni, 06] Baldoni, M., Baroglio, C., Brunkhorst, I., Henze, N., Marengo, E. Patti, V.: A Personalization Service for Curriculum Planning. In proceedings of 14th Workshop on Adaptivity and User Modeling in Interactive Systems, Hildesheim, 2006.

[Bizer, 09] Bizer, C., Heath, T., Berners-Lee, T.: Linked data - The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[Bodenreider, 04] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 32, 267-270, 2004.

[Breslin, 07] Breslin J, Decker S. : The future of social networks on the Internet. The need for semantics. IEEE Internet Computing. 11, 87-90, 2007

[Jonassen, 04] Jonassen, D., Churchill, D., Is there a learning orientation in learning objects? International Journal on E-learning 3(2), 32-41, 2004.

[Dertouzos, 97] Dertouzos, M.: What will be. London, UK: Judy Piatkus Ltd. (pp. 175-189), 1997.

[Devedzic, 07] Devedzic V., Jovanovic J., Gasevic D.: The pragmatics of current e-learning standards, IEEE Internet Computing, 11(3), 19-27, 2007.

[Dietze, 08] Dietze, S., Gugliotta, A., Domingue, J.: Supporting Interoperability and Context-Awareness in E-Learning through Situation-driven Learning Processes, Special Issue on Web-based Learning of International Journal of Distance Education Technologies (JDET), IGI Global, 2008.

[Dietze, 12] 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. Riva del Garda (Trento), Italy, 2012.

[Dietze, 11] Dietze, S., Yu, H. Q., Pedrinaci, C., Liu, D., Domingue, J.: SmartLink: a Web-based editor and search environment for Linked Services, In Proceedings of 8th Extended Semantic Web Conference. Heraklion, Greece, 2011.

[Engeström, 05] Engeström J.: Why some social network services work and others don't. The case for object-centered sociality. 2005; http://www.zengestrom.com/blog/2005/04/ why_some_social.html)

[Faro, 11] Faro, A., Giordano, D., Maiorana, F.: Mining massive datasets by an unsupervised parallel clustering on a GRID: Novel algorithms and case study. Future Generation Computer Systems, 27(6), 711-724, 2011.

[Giordano, 09] Giordano, D., Faro, A., Maiorana, F., Pino, C., Spampinato, C.: Feeding back learning resources repurposing patterns into the "information loop": opportunities and challenges. In the Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine. Larnaca, Cyprus, 2009.

[Grimnes, 08] Grimnes, G. A., Edwards, P., Preece, A.: Instance based clustering of semantic web resources. In S. Bechhofer et al. (Eds). The semantic web: research and applications. Proceedings of the 5th European semantic web conference. Tenerife, Canary Islands, Spain (pp. 303-317). LNCS 502: Springer-Verlag, 2008.

[Heath, 11] Heath, T. and Bizer, C. Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

[Hendrix, 12] Hendrix, M., Protopsaltis, A., Dunwell, I., de Freitas, S., Petridis, P., Arnab, S., Dovrolis, N., Kaldoudi, E., Taibi, D., Dietze, S., Mitsopoulou, E., Spachos, D., Bamidis, P., Technical Evaluation of The mEducator 3.0 Linked Data-based Environment for Sharing Medical Educational Resources, 2nd International Workshop on Learning and Education with the Web of Data, Lyon, France; 04/2012

[Henze, 04] Henze, N., Dolog, P., Nejdl, W.: Reasoning and Ontologies for Personalized E-Learning. Educational Technology & Society, 7(4), 82-97, 2004.

[Henze, 06] Henze, N.: Personalized E-Learning in the Semantic Web. International Journal of Emerging Technologies in Learning (iJET), 1(1), 2006.

[IEEE,02] IEEE, IEEE Standard for Learning Object Metadata, IEEE Std 1484.12.1-2002, 2002.

[Jonassen, 06] Jonassen, D., Churchill, D. (2006). Where is the Learning in Learning Objects?. *International Journal on E-Learning*. Chesapeake, VA: AACE

[Kaldoudi, 11a] Kaldoudi, E., Dovrolis, N., Giordano, D., Dietze, S.: Educational resources as social objects in semantic social networks. In the Proceedings of the Linked Learning 2011: 1st International Workshop on eLearning Approaches for the Linked Data Age, Herakleio, Greece, 2011.

[Kaldoudi, 11b] Kaldoudi, E., Dovrolis, N., Dietze S.: Information organization on the Internet based on heterogeneous social networks. In Proceedings of the 29th ACM International Conference on Design of Communication, (pp. 107-114). Pisa, Italy, 2011.

[Kobilarov, 08] Kobilarov, G., Dickinson, I.: Humboldt: Exploring Linked Data, In the Proceedings of Linked Data on the Web (LDW'08), Beijing, China, 2008.

[Kohonen, 95] Kohonen, T.: Self-organizing maps. Springer Series in Information Sciences, Vol 30, Springer 1995.

[Knorr-Cetina, 97] Knorr-Cetina K.: Sociality with objects: social relations in postsocial knowledge societies. Theory, Culture & Society. 1997; 14(4): 1–30.

[Lenz, 07] Lenz, R., Beyer, M., Kuhn, K.A.: Semantic Integration in Healthcare Networks, International Journal of Medical Infomatics. 76, 201-207, 2007.

[Marchionini, 06] Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM, 49(4):41-46, 2006.

[Mitsopoulou, 11] Mitsopoulo, .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, in the Proceedings of Linked Learning 2011: 1st Int. Workshop on eLearning Approaches for the Linked Data Age, Herakleio, Greece, 2011.

[Nilsson, 03] Nilsson, M., Palmer, M., Brase J.: The LOM RDF binding—principles and implementation. In Proceedings of the 3rd Annual ARIADNE Conference, Leuven, Belgium, 2003.

[Noy, 09] Noy, N. F., Shah, N. H., Whetzel, P.L., Dai B., Dorf, M., Griffith N., Jonquet, C., Rubin D. L., Storey, M.A., Chute C. G., Musen M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research, 37(2), 170-173, 2009.

[O'Reilly, 07] O'Reilly T. : What is Web 2.0: Design patterns and business models for the next generation of software. Communications & Strategies, 1(65), 17-37, 2007.

[Pedrinaci, 10] Pedrinaci, C., Liu, D., Maleshkova, M., Lambert, D., Kopecky, J., Domingue, J.: iServe: a Linked Services Publishing Platform, In the Proceedings of the Workshop on Ontology Repositories and Editors for the Semantic Web at 7th Extended Semantic Web Conference, 2010.

[Prakash, 09] Prakash, L. S., Saini, D. K., Kutti. N. S.: Integrating EduLearn learning content management system (LCMS) with cooperating learning object repositories (LORs) in a peer to peer (P2P) architectural framework. SIGSOFT Software Engineering Notes, 34(3), 1-7, 2009.

[de Santiago, 10] de Santiago, R., Raabe, A. L. A.: Architecture for learning objects sharing among learning institutions-LOP2P. IEEE Transactions on Learning Technologies, 3(2), 91-95, 2010.

[Konstantinidis, 09] Konstantinidis, S., Kaldoudi, E., Bamidis, P.: Enabling content sharing in contemporary medical education: a review of technical standards. The Journal on Information Technology in Healthcare, 7(6), 363-375, 2009.

[Robertson, 04] Robertson, S.: Understanding Inverse Document Frequency: On theoretical arguments for IDF. Journal of Documentation 60 no. 5, pp 503–520, 2004.

[Sahlgren, 05] Sahlgren, M.: An introduction to random indexing. In Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE), Copenhagen, Denmark, 2005.

[Schmidt, 04] Schmidt, A., Winterhalter, C.: User context aware delivery of e-learning material: approach and architecture, Journal of Universal Computer Science, 10(1), 38-46, 2004.

[Sheth, 08] Sheth, A. P., Gomadam, K., Ranabahu, A.: Semantics enhanced services: Meteor-s, SAWSDL and SA-REST. IEEE Data Engineering Bulletin, 31(3), 8-12, 2008.

[Simon, 04] Simon, B., Dolog., P., Miklós, Z., Olmedilla, D., Sintek, M.: Conceptualizing smart spaces for learning. Journal of Interactive Media in Education, 2004.

[Simon, 05] Simon, B., Massart, D., Assche, F., Ternier, S., Duval, E., Brantner, S., Olmedilla, D., Miklos. Z.: A simple query interface for interoperable learning repositories. In Proceedings of the 1st Workshop On Interoperability of Web-Based Educational Systems, 2005.

[Vitvar, 08] Vitvar, T., Kopecky, J., Viskova, J., & Fensel, D.: Wsmo-lite annotations for web services. In Hauswirth, M.; Koubarakis, M.; Bechhofer, S. (Eds.). In Proceedings of the 5th European SemanticWeb Conference, LNCS. Berlin, Heidelberg: Springer Verslag, 2008.

[Yu, 11] Yu, H. Q., Dietze, S., Li, N., Pedrinaci, C., Taibi, D., Dovrolis, N., Stefanut, T., Kaldoudi, E., Domingue, J.: Linked data-driven & service-oriented architecture for sharing educational resources. In Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age, CEUR-717, 2011.