Interlinking educational Resources and the Web of Data – a Survey of Challenges and Approaches

Stefan Dietze[1, 2], Salvador Sanchez-Alonso[3], Hannes Ebner[4],
Hong Qing Yu[2], Daniela Giordano[5], Ivana Marenzi[1], Bernardo Pereira Nunes[1, 6]

[1] L3S Research Center, Leibniz University Hannover, Appelstr. 9a, 30167 Hannover, Germany
{dietze; marenzi}@l3s.de

[2] Knowledge Media Institute, The Open University, MK7 6AA, Milton Keynes, UK
h.q.yu@open.ac.uk

[3] University of Alcala, Information Engineering research unit, Computer Science Department. Edificio Politécnico O246, Ctra. De Meco s.n., Alcalá de Henares, Madrid, Spain
salvador.sanchez@uah.es

[4]Royal Institute of Technology, School of Computer Science and Communication, Lindstedtsvägen 3, 10044 Stockholm, Sweden
hebner@csc.kth.se

[5] University of Catania, Dipartimento di Ingegneria Elettrica, Elettronica e Informatica, Viale A. Doria 6, 95125, Catania, Italy
dgiordan@diit.unict.it

[6] Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, RJ, 22543-900, Brazil
bnunes@inf.puc-rio.br

Abstract. Research in the area of 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, or interface mechanisms. 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 (e.g. RDF, SPARQL). To this end, it is obvious that the application of Linked Data principles offers a large potential to solve interoperability issues in the field of TEL. In this paper, we survey approaches aimed towards our vision of Linked Education, i.e. education which exploits educational Web data. This particularly considers the exploitation of the wealth of already existing TEL data on the Web by allowing its exposure as Linked Data and by taking into account automated enrichment and interlinking techniques to provide rich and well-interlinked data for the educational domain.

Keywords: Linked Data, Education, Semantic Web, SOA, E-Learning, Technology-enhanced learning (TEL), Web Data, Open Educational Resources.

1 Introduction

Throughout the last decade, research in the field of technology-enhanced learning (TEL) has focused fundamentally on enabling interoperability and reuse of learning resources and data. That has led to a fragmented landscape of competing metadata schemas, i.e., general-purpose ones such as Dublin Core[1] or schemas specific to the educational field, like IEEE Learning Object Metadata (LOM) (IEEE, 2002) or ADL SCORM[2] but also interface mechanisms such as OAI-PMH[3] or SQI[4]. These technologies are exploited by educational resource repository providers to support interoperability. To this end, although a vast amount of educational content and data is shared on the Web in an open way, the integration process is still costly as different learning repositories are isolated from each other and based on different implementation standards (De Santiago and Raabe, 2010).

In the past years, TEL research has already widely attempted to exploit Semantic Web (Berners-Lee *et al*., 2001) technologies in order to solve interoperability issues. However, while the Linked Data (LD) (Bizer *et al*., 2008; 2009) approach has widely established itself as the de-facto standard for sharing data on the Semantic Web, it is still not widely adopted by the TEL community. Linked Data is based on a set of well-established principles and (W3C) standards, e.g. RDF, SPARQL (World Wide Web Consortium*.,* 2008) and use of URIs, and aims at facilitating Web-scale data interoperability. Despite the fact that the LD approach has produced an ever growing amount of data sets, schemas and tools available on the Web, its take-up in the area of TEL is still very limited. Thus, Linked Data opens up opportunities to substantially alleviate the interoperability issues in the field, some of which were addressed above.

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 for the TEL field is to (a) start adopting LD principles and vocabularies while (b) leveraging on existing educational data available on the Web by non-LD compliant means. Following such an approach, four major research challenges need to be taken into consideration to ensure Web-scale interoperability:

(C1) *Integrating distributed data from heterogeneous educational repositories*: educational data and content is usually exposed by heterogeneous services/APIs such as OAI-PMH or SQI. Therefore, interoperability is limited and Web-scale sharing of resources is not widely supported yet (Prakash *et al*., 2009).

(C2) *Dealing with continuous change*: in highly distributed Web-based environments, frequent changes occur to available Web APIs. That is, services as well as repositories are usually added, modified or removed regularly.

(C3) *Metadata mediation and transformation*: educational resources and the services exposing those resources are usually described by using distinct, often XML-based schemas and by making use of largely unstructured text and heterogeneous taxonomies. Therefore, schema and data transformation (into

[1] http://dublincore.org/documents/dces/

[2] Advanced Distributed Learning (ADL) SCORM: http://www.adlnet.org

[3] Open Archives Protocol for Metadata Harvesting http://www.openarchives.org/OAI/openarchivesprotocol.html

[4] Simple Query Interface: http://www.cen-ltso.net/main.aspx?put=859

RDF) and mapping are important requirements in order to leverage on already existing TEL data.

(C4) *Enrichment and interlinking of unstructured metadata*: existing educational resource metadata is usually provided based on informal and poorly structured data. That is, 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 LD vocabularies and datasets on the Web.

In this paper we provide a survey of general approaches which serve as building blocks towards *Linked Education*[5], i.e. educational processes enabled by the vast interconnected cloud of Web data. Therefore we identify approaches which address the above challenges by following and supporting the below principles:

(P1) **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' description and resources are exposed in a standardized and accessible way. This serves as overall principle for all subsequent activities.

(P2) **Services integration:** 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 (see P1).

(P3) **Schema matching:** metadata retrieved from heterogeneous Web repositories, for instance IEEE LOM resource metadata, needs to be automatically lifted into RDF and mapped with competing metadata schemas and exposed as Linked Data accessible via de-referenceable URIs.

(P4) **Data interlinking, clustering and enrichment:** Automated enrichment and clustering mechanisms are exploited in order to interlink data produced by (P3) with existing datasets as part of the LD cloud.

The remaining sections of the paper start with a general overview of related research, followed by a description of key challenges and research areas in Section 3. In Section 4 we survey related work in the field of educational services and APIs while we assess educational data integration techniques and standards in Section 5. We summarize all surveyed technologies in Section 6 and provide an assessment of how particular technologies contribute to the challenges described above.

2 General overview on related research

To facilitate a better understanding of the overall challenges, from a research as well as a pragmatic perspective, we provide an initial overview on related work in this section before providing a more elaborate assessment of related technologies in the subsequent sections.

[5] http://linkededucation.org: an open platform to share results focused on educational LD. Long-term goal is to establish links and unified APIs and endpoints to educational datasets.

Web-scale search of educational resources faces a heterogeneous landscape of Web APIs of individual repositories. For instance, the PubMed[6] repository provides an OAI-PMH-based service where response messages are based on XML in OAI-DC (OAI Dublin Core) while other repositories offer JSON-based feeds or SPARQL endpoints, such as the Linked Data store from The Open University (UK)[7]. In addition, current metadata stores largely use XML and relational databases and consist largely of poorly structured text and lack formal semantic. That leads also to largely ambiguous descriptions which are hard to interpret and process at the machine-level.

Services operating on educational repositories are very dynamic, in that APIs appear and are removed from the Web frequently and might change behaviors and interfaces according to new requirements. Therefore, it is crucial to aim at shielding the underlying heterogeneity and minimize disturbance of upper layers (e.g. educational applications). Therefore, facilitating easy-to-use service representations based on standard service vocabularies, e.g. SAWSDL (Sheth *et al.*, 2008) and WSMO-Lite (Kopecky *et al*., 2008; Vitvar *et al*., 2008) is an important requirement to allow service providers and consumers to interact. In this paper, we are applying LD technologies to both (a) educational service and APIs and (b) educational data in order to facilitate data as well as services interoperability. The main principles of LD (Bizer *et al*., 2008) imply the use of (dereferenceable) URIs to identify things, the use of RDF and SPARQL for data representation and interaction and the interlinking of datasets. The above principles have proven largely successful throughout the past years, leading to an ever increasing amount of LD-compliant schemas and data-sets[8] as well as general-purpose tools and APIs.

Several efforts were already made to improve interoperability in the field of education, e.g. by exploiting semantic technologies. For instance, efforts have been made on providing an IEEE LOM-RDF binding[9] and its mapping into the Dublin Core Abstract model, but this early work was (a) discontinued and (b) only focused on the binding aspect rather than further working towards a Linked Data-compliant approach, e.g. by reusing elements of established Linked Data schemas or linking vocabularies. A peer-to-peer (P2P) architecture (LOP2P) for sharing educational resources among different learning institutions is proposed in (De Santiago and Raabe, 2010). LOP2P aims at creating course material by using shared educational resource repositories based on a particular LOP2P plugin. A similar P2P architecture has also been proposed in the EduLearn project (Prakash *et al*., 2009). Meanwhile, Simple Query Interface (SQI) is introduced in (Ternier *et al.*, 2006) designed to query different learning repositories using a comment query language. However, query format and result format have to be agreed among different repository providers before using the query functionalities, which means that a wrapper service is required to ensure compliancy of all involved repositories with the agreed format. These approaches are sharing a number of disadvantages. For instance, instead of accepting the heterogeneous landscape of the Web, all approaches impose either a common

[6] http://www.ncbi.nlm.nih.gov/pubmed/
[7] http://data.open.ac.uk/
[8] http://richard.cyganiak.de/2007/10/lod/
[9] http://dublincore.org/educationwiki/DCMIIEEELTSCTaskforce/RDFPAR

schema or interface approach on the underlying stores. Also, mediation is based on syntactic matching, which does not deal well with ambiguities.

The work described in (Schmidt and Winterhalter, 2004) and (Schmidt, 2005) utilizes Semantic Web as well as Web service technologies to enable adaptation to different learning contexts by introducing a matching mechanism to map between a specific context and available learning data. However, this work neither considers approaches for automatic service discovery nor it is based on common standards. Also, mediation between different metadata standards is not supported. (Dietze *et al.*, 2008) follows a similar approach but has scalability issues as it is fundamentally based on a single shared ontology. These issues apply as well to the idea of "Smart Spaces" (Simon *et al*., 2004) for learning. The work in (Baldoni *et al*., 2006) follows the idea of using a dedicated personalization Web service that makes use of semantic learning object descriptions to identify and provide appropriate learning content. Neither is the integration of several distributed learning services within the scope of this research, nor is the allocation of services at runtime. Further related research on (Henze, 2006) and (Henze *et al*., 2004) allows a mediation between different services based on a so-called "connector service".

From a more pragmatic angle, educational institutions started to expose their data based on Linked Data principles, such as The Open University (UK)[10], the National Research Council (CNR, Italy)[11] or Southampton University (UK)[12]. However, while that 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 3rd party data is not yet within the primary scope.

3 Towards Linked Education: research challenges and areas

In this section, we provide an overview of a general-purpose approach which aims at (i) integrating heterogeneous educational Web resources and (ii) exposing its metadata as well-structured and interlinked Linked Data. Our overall proposed architecture includes three layers: *Educational (Web) data and service layer*, *Educational data and service integration layer* and *Educational application and presentation layer* that are shown in Figure 1.

- The *Educational (Web) data and service layer* consists of available educational Web services and data, such as metadata of existing educational objects provided by open public educational repositories, such as PubMed[13] or OpenLearn[14].
- The *Educational data and service integration layer* is fundamentally based on exploiting Linked Data principles to annotate and interlink educational services and data. The *Educational application and presentation layer* uses the APIs

[10] http://data.open.ac.uk
[11] http://data.cnr.it
[12] http://data.southampton.ac.uk/
[13] http://www.pubmed.gov
[14] http://www.open.ac.uk/openlearn

provided by the educational data & services integration layer to interact with underlying data & services and provides an interface to end-users.

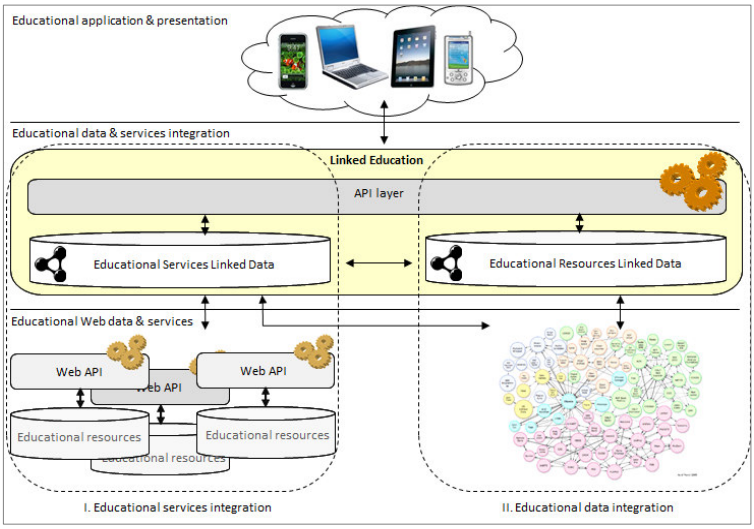

Fig. 1. Educational Web data integration - overview.

To enable a wide integration of disparate Web data, two main steps are required which are both facilitated by Linked Data technologies:

Step I. Educational Services Integration
(facilitated by *Educational Services Linked Data* on the left)
Step II. Educational Data Integration
(facilitated by *Educational Resources Linked Data* on the right)

As laid out in Section 1, integration of educational data and content needs to consider two challenges: integration at the repository-level facilitated by repository-specific APIs and integration at the (meta)data-level. *Step I* aims at integrating educational services and APIs in order to facilitate repository-level integration. To this end, it is concerned with resolving heterogeneities between individual API standards (e.g. SOAP-based services vs. RESTful 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, we exploit Linked Data principles to annotate individual APIs in terms of their interfaces, capabilities and non-functional properties (*Educational Services Linked Data*). That allows to automatically discover and execute APIs for a given educational purpose (for instance, to retrieve educational metadata for a given subject and language) while resolving heterogeneities between individual API responses (as detailed in the following Section). All educational data retrieved in *Step I* will be transformed from their native formats into RDF.

Step II, deals with the actual integration of heterogeneous educational data as retrieved by Step I by exposing all retrieved educational (RDF) data as well-interlinked Linked Data. As starting point, all generated RDF is stored in a dedicated, public RDF store (*Educational Resources Linked Data*) which supports two main

purposes: exposing existing educational (non-RDF) data in a LD-compliant way and allowing content/data providers to publish new educational resource metadata. To enrich and interlink the educational data, two approaches are being followed:

1. Automated interlinking of datasets
2. Automated clustering and classification.

While exposing educational data as RDF is one substantial requirement to follow LD principles, mere transformation of data does not improve its quality. For instance, largely unstructured metadata descriptions retrieved in Step 1 as part of XML-based metadata descriptions do not automatically benefit from a mere transformation into RDF. Thus, it is even more challenging to enrich such unstructured descriptions by automated data enrichment techniques to establish links with established vocabularies available on the LD cloud. Enrichment takes advantage of available APIs such as the ones provided by DBPedia Spotlight[15] or Bioportal[16], which allow access to a vast number of established taxonomies and vocabularies. That way, unstructured free text is enriched with unique URIs of structured LD entities to allow not only further reasoning on related concepts but also enables users to query for resources by using well-defined concepts and terms. In addition, automated clustering and classification mechanisms are exploited in order to enable data and resource classification across previously disconnected repositories.

4 Educational services integration: Web APIs and interfaces

During the last 15 years, portals offering access to educational materials in digital format – often called learning object repositories as an "umbrella" term – have grown to the point of becoming an essential component for eLearning. If we look at these portals individually, all provide searching, browsing and navigation services across their collections, but in many cases the reach of these services is limited to the collection of each particular repository, in what some authors have called "isolated silos of information" (Ochoa and Duval, 2009).

A recent tendency is to provide capabilities to extend searching capabilities to resources or metadata distributed across different repositories (Klemke *et al.,* 2010). Technically, this can be mainly achieved through two mechanisms: harvesting and distributed search. In the harvesting model, a central location gives access to learning resources from a number of different sources by collecting the metadata into a central location. In the distributed search model, a query is spread out over several repositories and the individual results put together and eventually ranked according to some criteria. In what follows of this section we will further explore the technologies that make these two models possible.

15 http://dbpedia.org/spotlight
16 http://www.bioontology.org/wiki/index.php/BioPortal_REST_services

4.1. Services and interfaces to share educational Web repositories

This section assesses common Web interfaces used by educational data stores to expose and share data and metadata.

SQI

Federated search is a term that defines the capability of a repository to search beyond its own borders; e.g. MERLOT[17] allows searching simultaneously in 20 partner collections and digital libraries. By spreading a query across many collections at once, aggregating search outcomes and providing a single list of results, repositories offer a very valuable service to their users, who no longer need to repeatedly visit and query each and every data source in the federation. This information retrieval technology allows users to carry out a more efficient search process while obtaining higher quality and more relevant results, fostering at the same time a "plug-in" model: with each new repository added to the federation, more content becomes available to its community of users.

In this context, it is relevant to devote some time to the Simple Query Interface (SQI), a specification for querying learning object repositories. Supported and promoted by the CEN/ISSS Learning Technologies Workshop, SQI was made public in 2004 as a purposefully simple specification, neutral in terms of results format and query languages, and allowing to combine searches in heterogeneous repositories. However, SQI "does not directly contribute to overcome the differences of the various paradigms in metadata management" having been designed instead to "become an independent specification for all open educational repositories" (Simon *et al.*, 2005).

What SQI defines is a common communication interface between repositories, a list of methods (query methods, configuration methods and session management methods) that all repositories must make available before they can receive and answer queries from external systems. In this regard, a SQI compliant repository must wrap their internal query language, metadata schema and results format and expose a SQI interface to any external system that might like to query it. Some interesting characteristics of SQI are the following (CEN, 2005):

- SQI is agnostic on Query Language and Results Format
- It can be deployed in both synchronous and asynchronous search scenarios
- SQI methods can either perform an action or carry out query, but not both (command-query separation principle)
- Commands are simple and extensible
- SQI separates queries from session management by using a very simple approach: no queries can take place if no session has been established.

In 2009, a good number of repositories – e.g. Merlot, OERCommons[18] or the then popular EdNA Online[19] – were implementing SQI with an acceptable degree of compliance (Hilera *et al.*, 2009). Today, however, ambitious projects and

17 http://www.merlot.org
18 http://www.oercommons.org/
19 http://www.edna.edu.au/

architectures such as Ariadne[20], where the number of repositories to query is high and in constant growth, are moving to different schemas of operation mainly because of SQI models shortcomings such as the demand for the implementation and agreement of a query language specific to each client-server pair (Ternier *et al*., 2008).

OAI-PMH

Currently the basis of most interoperability efforts in the learning object repositories field, OAI-PMH is a protocol for transferring over the Web metadata about any material stored in electronic form. Compared with similar initiatives such as Z39.50 (which addresses issues such as session management, management of result sets and specification of predicates to filter out results), OAI was intentionally designed simple to reduce implementation complexity and therefore costs, facilitating its adoption. Nonetheless, this new protocol was designed for transferring large amounts of metadata, and thus provides a reasonable solution for clients that need to aggregate or index metadata (Haslhofer and Schandl, 2008).

The OAI-PMH model is called "metadata harvesting", a solution that enables providers of information resources (e.g. learning object repositories) expose their metadata via an interface that can be used later as the basis for the development of value-added services. As mentioned in the introduction to this section, the protocol allows ingesting metadata into a central metadata repository, where new services operating on the metadata from the many harvested sources can be implemented and offered to the community of users.

OAI-PMH uses HTTP transactions to issue questions to a client (a metadata collector service) and get answers from a metadata server. A client may e.g. request a server to send metadata according to certain criteria such as date of creation of the data and, in response, the server would return a recordset in XML format that should include identifiers (e.g. URLs) of the objects described in each record. In order to formulate these questions, OAI-PMH offers a catalogue of "verbs" that can be used to request the information that is needed according to the circumstances. An "Identify" request, for instance, retrieves administrative metadata about a repository such as name or owner, whereas "GetRecord" allows fetching the metadata record for a certain resource.

The protocol supports multiple formats for expressing metadata. However, all servers implementing OAI-PMH must support the Dublin Core metadata element set as the common format, a decision deriving from the original need to provide a shared and widely disseminated format among the community of potential users (Lagoze and Van de Sompel, 2001; 2003). In learning objects repositories, this flexibility to implement different metadata models has allowed to implement systems using IEEE LOM application profiles as the basis for the description of their metadata (e.g. Organic.Edunet).

Currently, thousands of institutions worldwide maintain OAI-PMH repositories and data sources (http://www.openarchives.org/Register/BrowseSites), and many commercial and repository software systems exist (e.g. DSpace, Fedora or ePrints). However, it is not exempt from drawbacks as Haslhofer and Schandl (2008) pointed out: resources are not accessible through de-referenceable URIs and the selective

20 http://www.ariadne-eu.org/

access to metadata is restricted. The same authors expose OAI-PMH Metadata as Linked Data via a wrapper server that provides interfaces for external access (e.g. a SPARQL endpoint).

In learning repositories, the use of OAI-PMH can be used as the basis for aggregation – i.e. harvesting different sources of information from several containers – thus enhancing the repository value and providing better services to end-users. Once compiled, aggregations can act (a) as the basis for metadata generation, (b) for the provision of new services working on the new metadata and, very often, (c) for the re-exposure of the metadata via OAI-PMH for further aggregation. A good example of this is the Organic.Edunet federation of repositories. Organic.Edunet harvests SCAM repositories for learning resources on organic agriculture, stores them locally and provides metadata and advanced search and retrieval mechanisms based on this new metadata but, at the same time, it is harvested by Ariadne as part of the so-called "Ariadne infrastructure" playing the role of just another repository in the Ariadne federation.

SPARQL endpoints

SPARQL endpoints are services that enable users to query a knowledge base via the SPARQL language. In some way, they can be considered machine-oriented interfaces to online databases. RDF triplestores, i.e. databases providing persistent storage and access to RDF graphs, usually provide SPARQL endpoints. In this sense, some of the most widely used triplestore systems (e.g. Virtuoso[21], AllegroGraph[22], Joseki[23], Sesame[24] or Mulgara[25] just to name a few), implement some form of SPARQL endpoint:

- AllegroGraph implements the 4.3.3 HTTP Protocol, a super-set of the Sesame 2.0 HTTP protocol and the W3C SPARQL protocol that allows AllegroGraph servers to expose several data catalogs, each containing any number of repositories (triplestores).
- Mulgara provides a SPARQL query parser and query engine, including some extensions and purpose-built modifications, e.g. it only allows results to be ordered by variables.
- Openlink's Virtuoso query service provides a SPARQL endpoint allowing to perform SPARQL queries as well as uploading of data over HTTP. This query service extends the standard protocol to provide additional features, e.g. support for multiple output formats.
- Sesame native also integrates a web server and SPARQL endpoint. Sesame is different in the sense that all other triplestores mentioned above can be used through the Sesame API.
- Although not a native triplestore, Drupal[26] version 7 offers a SPARQL module extension which includes the core SPARQL API functionality. This module,

21 http://virtuoso.openlinksw.com
22 http://www.franz.com/agraph/allegrograph/
23 http://www.joseki.org/
24 http:// www.openrdf.org/
25 http://www.mulgara.org/
26 http://drupal.org/

which requires the RDF Drupal module to expose the data contained in the CMS as RDF, includes the SPARQL registry and the SPARQL endpoint modules.
- HP Jena[27] includes ARQ, an implementation of the SPARQL query language offering "legal" SPARQL syntax as well as some extensions such as GROUP-BY or SELECT expressions.
- Joseki, another Hewlett Packard effort, also supports the SPARQL Protocol through a HTTP (GET and POST) implementation.

As Cheung *et al.* (2009) commented on his work, SPARQL helps solving interoperability problems derived from the underlying different technologies of each triplestore. Thus, SPARQL allows datasets in each triplestore to be accessed via standard SPARQL queries issued by clients to a common SPARQL endpoint service, an approach which allows creating cross-links at programming level (Cheung *et al.,* 2009).

From a linked data point of view, the recent availability of frontend tools for SPARQL endpoints such as Pubby is remarkable, as they allow creating linked data interfaces to SPARQL endpoints. In any case, an increasing number of triplestores (e.g. Virtuoso) provide native linked data exposure as part of their functionality. We would like to finally point out the importance and widespread use of SPARQL by mentioning that DBpedia – according to many one of the more famous parts of the Linked Data efforts – provides a public SPARQL endpoint which enables users to query the RDF datasource with SPARQL queries.

Other technologies (including proprietary APIs)

Apart from the previously described technologies, others such as OKI OSID (Ternier *et al.,* 2006) or SRU/SRW (Morgan, 2004) co-exist. At the same time, many repositories offer proprietary interfaces to access the data stored in their databases. Here we mention a few cases that can serve as an example of what there is today:

- Library of Congress' SRU/SRW pair of protocols are an evolution over the old Z39.50, which they replace by HTML. SRU (Search/Retrieve via URL) utilizes CQL (Contextual Query Language) to return results in XML and enables URL query strings, while SRW (Search Retrieve Web Service) is a complement of SRU which provides a SOAP interface to SRU queries.
- OKI OSID (Open Services Interface Definition) is an abstraction including browsing, searching and other access to a repository implemented by many resource collections such as Connexions[28], Merlot or MIT OpenCourseWare[29].
- Merlot Web services include different search syntaxes which allow direct search capability into the MERLOT collection from a Web-based application. The repository offers a "basic simple search" Web Service available to any MERLOT members, as well as a suite of more elaborated web services for selected partners.
- The Connexions repository offers a proprietary REST API which uses JavaScript and XML-RCP protocols to return results in XML.

27 http://jena.sourceforge.net
28 http://cnx.org/
29 http://ocw.mit.edu

- Edna[30] APIs (HTML, XML and RSS formats available) enabled Google-like searches with limited query possibilities. All Edna services discontinued from September 2011.
- OpenCourseFinder[31] provides two APIs[32] to access the information published by selected universities in OCW format: The OpenCourseWare Search API (which allows integrating OCW Search results into external applications) and the OCW Search Meta Data API (which gives access to the metadata OCW Search tracks for all the courses in their collection).

This list is not at all intended to be an exhaustive survey of all available access technologies, as we consider that such thing is indeed beyond the scope of this paper, but we hope it can somehow serve as a general overview of the current situation.

4.2. Integration of heterogeneous educational services

This section reports on several prominent approaches to integrate distributed stores into federated applications, e.g. Ariadne, Luisa and other Semantic Web or Semantic Web Services-based approaches.

The **LUISA** project created a repository framework called LOM-R, later evolved into Ont-Space[33], offering semantic search functionalities over the metadata stored in an ontology language (WSML in LOM-R and OWL in the latest implementation Ont-Space). This project demonstrated that metadata records from heterogeneous sources can be "translated" into a common ontological format and later stored in a semantic repository offering uniform capabilities. Its main result, the Ont-Space framework, has been used as a basic part in the architecture of subsequent efforts such as Organic.Edunet[34] and VOA^{3}R project[35].

The **Organic.Edunet** portal is a central point of access to the resources stored by a federation of repositories hosting thousands of digital learning resources about organic agriculture and agroecology. Its two-level architecture relies on OAI-PMH for the portal to harvest learning objects metadata from the repositories, eventually enrich them and exposing them to its users. Metadata in Organic.Edunet must comply with the IEEE Learning Object Metadata (LOM) Application Profile (AP) for Organic.Edunet learning resources. As some sources were not part of the content providing effort carried out during the project timeline (2007-2010), many metadata records need completion, enrichment or just validation before the educational resources they refer to are made available through the portal. But Organic.Edunet is not only a harvester, as it also provides several interfaces to expose the metadata (e.g. SQI and OAI-PMH), and thus is harvested by wider-scope federations such as Ariadne. The portal is currently under a redesign process to export linked data (Sicilia *et al.,* 2011) as it will be detailed in Section 5.

30 http://edna.edu.au
31 http://www.ocwfinder.org
32 http://www.ocwsearch.com/api/
33 http://code.google.com/p/ont-space/
34 http://www.organicedunet.org/
35 http://voa3r.eu/

The **Ariadne** infrastructure is another good example of an aggregated system based on the use of OAI-PMH technologies. Ariadne currently provides access to several hundreds of thousands of learning resources from repositories and collections around the world operating under a dual model: hosting repositories of institutions without a specific infrastructure of their own, and locally storing metadata records from other institutions maintaining and hosting their own repositories. On top of the repository, the Ariadne infrastructure provides a Simple Query Interface (SQI) which basically acts as a gateway to underlying specifications such as SRU/SRW or OKI OSIDs.

An older example – an ancestor in some way of many of current integration initiatives –, would be the **eduSource** Canada project, now discontinued. Intended as a collaboration among Canadian public and private sectors, its main aim was to draft and test an open network of interoperable learning object repositories (Downes, 2004). This approach was based on the use of the eduSource Communications Layer protocol (ECL), which implemented the IMS DRI model (IMS DRI, 2003). Edusource was also pioneer in the research in Semantic Web technologies and their application to learning object repositories.

The **mEducator project**[36] aims is to analyze the use of existing standards and reference models in the educational field by providing and evaluating reference implementations aimed at discovery, retrieval, sharing and re-use of medical educational resources. Particular outcomes include a general architecture which follows the key principles described in this paper (Dietze *et al.,* 2012; Yu *et al.*, 2011), a Linked Data-compliant dataset (*mEducator Linked Educational Resources*[37]) describing educational resources according to a well-defined RDF schema (*mEducator Resource RDF schema*[38]). For integration of distributed educational data, a dynamic and Linked Data-based approach to services integration (Dietze *et al.*, 2011) is used.

More recently, the Asset Description Metadata Schema (ADMS)[39] has been developed as a common description layer across distinct repositories and metadata schemas. As such, it provides a metadata schema for describing particular information objects (i.e., assets), such as educational resources, as well as their source repositories. With ADMS' focus on assets and respositories themselves, as opposed to the API or interface used to expose the data, it complements existing service integration and schema mapping mechanisms (Section 4.3) by providing a common resource schema for alignment with heterogeneous ones.

4.3. Schema mapping

The work in (Dietze *et al.,* 2012) (Yu *et al.*, 2011) utilizes a semi-automated schema mapping approach in which mappings between schemas are defined in so-called lifting templates which are defined at design-time and applied at runtime to enable

[36] http://www.meducator.net
[37] http://thedatahub.org/dataset/meducator
[38] http://purl.org/meducator/ns
[39] http://www.w3.org/ns/adms

lifting of data from one schema to another. However, fully automated schema mapping approaches are partially exploited in the educational field as well. Fully automated approaches are on the one hand more scalable, but on the other hand, perform lower in terms of precision. In this section we briefly present schema and ontology matching techniques to cope with the lack of interoperability between the wide variety of learning repositories and property educational schemas available on the Web.

Schema and ontology matching rely on the task of automatically finding correspondences between elements or concepts between two or more data models (Massmann *et al*., 2011), aiming to create a unified view of data between different sources. Although Linked Data principles are straightforward, the conversion and integration of existing repositories developed in different data models and standards are a very hard and time-consuming task. In this manner, the great effort of the database community and others should also be applied to the TEL community in order to facilitate the integration of heterogeneous data, see (Doan and Halevy, 2005; Kalfoglou and Schorlemmer, 2003; Rahm and Bernstein, 2001; Shvaiko and Euzanat, 2005) for traditional surveys. In (Bernstein *et al*., 2011), the authors present future trends and a list of schema matching techniques.

COMA++[40] (Amuller *et al*., 2005; Do and Rahm, 2002) is a multi-strategy and graph-based approach able to combine multiple matching algorithms, reuse previous match mappings and support matching between different schemas and ontologies. A new version of this system is under development (Massmann *et al*., 2011), called COMA 3.0, and is expected to support ontology merge, data transformation and complex matching, which already is addressed by (Dhamankar *et al*., 2004; Nunes *et al*., 2011; Carvalho *et al*., 2008).

S-Match[41] (Giunchiglia *et al*., 2010) is a semantic matching framework for mapping lightweight ontologies. Their approach is based on removing ambiguities introduced by Natural Language through the use of Description Logic to relate nodes in different taxonomies. A similar approach is presented by (Raunich and Rahm, 2011). RiMOM (Li *et al*., 2009) is a framework responsible to find semantic matching between entities in different ontologies using a dynamic strategy to select and combine textual and structural metrics to generate the matching.

5 Educational data integration: enrichment, clustering and interlinking educational linked data

In this section, we describe the current landscape of educational metadata, in particular Open Educational Resources (OER) metadata, on the Web and discuss approaches which exploit Linked Data technologies for interlinking heterogeneous datasets.

[40] Binary version of COMA++ available at http://dbs.uni-leipzig.de/Research/coma_index.html
[41] S-Match is available at http://s-match.org/

5.1. Open Educational Resources & educational Linked Data: standards and repositories

Open Educational Resources (*OER*) are educational material freely available online. The wide availability of educational resources is a common objective for universities, libraries, archives and other knowledge-intensive institutions raising a number of issues, particularly with respect to Web-scale *metadata interoperability* or legal as well as *licensing aspects*. Several competing standards and educational metadata schemata have been proposed over time, including IEEE LTSC LOM[42] (*Learning Object Metadata*), one of the widest adopted, IMS[43], Ariadne, ISO/IEC MLR - ISO 19788[44] Metadata for Learning Resources (MLR) and Dublin Core (see also Koutsomitropoulos *et al*., 2010). The adoption of a sole metadata schema is usually not sufficient to efficiently characterize learning resources. As a solution to this problem, a number of taxonomies, vocabularies, policies, and guidelines (called *application profiles*) are defined (Duval *et al*., 2002). Some popular examples are: UK LOM Core[45], DC-Ed[46] and ADL SCORM.

Due to the diversity of exploited standards, existing *OER repositories offer very heterogeneous datasets*, differing with respect to schema, exploited vocabularies, and interface mechanisms. For example, MIT Open Courseware[47] (OCW), OpenLearn[48] is the UK Open University's contribution to the OER movement and it is a member of the MIT OCW Consortium. Video material from OpenLearn, distributed through iTunes U has reached more than 40 million downloads in less than 4 years[49]. One of the largest and diverse collections of OER can be found in the GLOBE[50] (Global Learning Objects Brokered Exchange) where jointly, nearly 1.2 million learning objects are shared. KOCW[51], LACLO[52] and OUJ[53] expose a single collection of metadata instances with a common provenance. Other repositories, such as ARIADNE, LRE[54], OER and LORNET[55] expose the result of the aggregation of several metadata collections that have different provenance.

Regarding the presence of *educational information in the linked data landscape*, two types of linked datasets need to be considered: (1) datasets directly related to educational material and institutions, including information from open educational repositories and data produced by universities; (2) datasets that can be used in teaching and learning scenarios, while not being directly published for this purpose.

---

42 http://ltsc.ieee.org/wg12/par1484-12-1.html
43 http://www.imsglobal.org/metadata/
44 http://www.iso.org/iso/
45 http://zope.cetis.ac.uk/profiles/uklomcore/
46 http://www.dublincore.org/documents/education-namespace/
47 http://ocw.mit.edu/index.htm
48 http://openlearn.open.ac.uk/
49 http://www.bbc.co.uk/news/education-15150319
50 http://globe-info.org/
51 http://www.koreabrand.net/
52 http://www.laclo.org/
53 http://www.ouj.ac.jp/eng/
54 http://lreforschools.eun.org/
55 http://www.lornet.org/

This second category includes for example datasets in the cultural heritage domain, such as the ones made available by the Europeana project[56], as well as by individual museums and libraries (such as the British Museum[57], who have made their collection available as linked data, representing more than 100 Million triples, or the Bibliothèque Nationale de France[58], who made available information about 30,000 books and 10,000 authors in RDF, representing around 2 Million triples). It also includes information related to research in particular domains, and the related publications (see PubMed[59] which covers more than 21 Million citations, in 800 Million triples), as well as general purpose information for example from Wikipedia (see DBPedia.org).

Regarding category (1), initiatives have emerged recently using linked data to expose, give access to and exploit public information for education. The Open University in the UK was the first education organization to create a linked data platform to expose information from across its departments, and that would usually sit in many different systems, behind many different interfaces (see http://data.open.ac.uk which includes around 5 Million triples about 3,000 audio-video resources, 700 courses, 300 qualifications, 100 Buildings, 13,000 people (Zablith *et al*., 2011a; 2011b). Many other institutions have since then announced similar platforms, including in the UK the University of Southampton (http://data.southampton.ac.uk) and the University of Oxford (http://data.ox.ac.uk). Outside the UK, several other universities and education institutions are joining the Web of Data, by publishing information of value to students, teachers and researchers with linked data. Noticeable initiatives include the Linked Open Data at University of Muenster[60] and the LODUM[61] project in Germany or the Norwegian University of Science and Technology exposing its library data as linked open data[62]. In addition, educational resources metadata has been exposed by the mEducator project (Mitsopoulou, *et al*., 2011; Dietze *et al.* 2012). A more thorough overview of educational Linked Data is offered by the Linked Education[63] platform.

5.2. Addressing OER metadata heterogeneity by adopting Linked Data principles

The problems connected to the heterogeneity of metadata can be addressed by converting the data into a format that allows for implementing the Linked Data principles (Bizer *et al.,* 2008). Most often this means that the data which is provided as part of RDBMS or in XML format – or, on occasion, in other formats – are converted into RDF. The data model of RDF is a natural choice as it allows for unique identification, interlinking to related data, as well as enrichment and

[56] http://www.europeana.eu/
[57] http://collection.britishmuseum.org/
[58] http://data.bnf.fr/
[59] http://www.ncbi.nlm.nih.gov/pubmed/ and http://thedatahub.org/dataset/bio2rdf-pubmed
[60] http://data.uni-muenster.de
[61] http://lodum.de
[62] http://openbiblio.net/2011/09/08/ntnu/
[63] http://linkededucation.org

contextualization. Therefore, general-purpose tools such as D2R[64], Virtuoso[65] and Triplify[66] are often used to convert proprietary datasets into RDF.

It is common to use DBpedia or other big datasets as "linking hubs" (Auer *et al*., 2007). One of the advantages of such an approach is that such datasets are commonly used by other datasets, which automatically leads to a plurality of indirect links. In the case of more specialized applications it is beneficial if domain specific datasets or ontologies can be found and linked to. This has been successfully demonstrated by specialized projects such as Linked Life Data[67] in the biomedical domain, Organic.Edunet[68] in organic agriculture and agroecology (Ebner *et al*., 2009), and mEducator[69] in medical education (Yu *et al*., 2011).

The approaches applied for creating links between datasets can be fully automatic, semi-automatic and fully manual. A lot of tasks required for interlinking and enhancing (enriching) metadata can be automated by analyzing textual content using Information Extraction (IE) and Natural Language Processing (NLP) techniques. Most commonly this includes the detection of sentences, named entities, and relationships, as well as disambiguation of named entities. However, quality control implies that the process has to be supervised at some point. The links can be created manually; alternatively the automatically detected links can be approved manually. NLP has its roots in machine learning which implies the use of learning algorithms which are trained on large textual corpora which eventually are domain-specific. Public services such as DBpedia Spotlight[70] and OpenCalais[71] offer NLP services relevant for linking data and also provide their output in RDF. In addition to these services which are ready to use, frameworks such as Apache Stanbol[72] can be easily integrated and provide solutions for the most common tasks involved in the creation of Linked Data, such as textual analysis and metadata extraction. The RESTful API allows for easy integration which should help projects dealing with metadata management using semantic technologies to hit the ground running.

Traditional ways of managing metadata often take a document-centric approach and use XML as it is an established standard for expressing information. Transformation of metadata into other formats requires a thorough mapping to be crafted, which often involves an analysis of the exact semantics of the involved standards. If such heterogeneous formats are to be transformed into Linked Data, good knowledge of existing standards is required as it is good practice to reuse established terms from other RDF-based standards (Nilsson, 2010) whenever possible. There are situations where the conceptual model of the origin data cannot be cleanly mapped to the RDF model and information may be lost. To avoid such situations, RDF should be considered as a basis for metadata interoperability (Nilsson,

---

64 http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/
65 http://virtuoso.openlinksw.com/
66 http://triplify.org/
67 http://www.linkedlifedata.com
68 http://www.organic-edunet.eu
69 http://www.meducator.net
70 http://dbpedia.org/spotlight
71 http://www.opencalais.com
72 http://incubator.apache.org/stanbol/

2010) – a common carrier – when adapting existing or creating new metadata standards.

The joint working group from IEEE LTSC and Dublin Core made an attempt to address heterogeneity of educational metadata by developing a mapping of IEEE LOM into the Dublin Core Abstract Model. This work resulted in a draft report in 2008, but the uptake has not been overwhelming. To date, the only known project to implement this draft[73] is the Organic.Edunet project, whose achieved goal it was to build a federation of learning repositories with material on organic agriculture and agroecology. The EntryStore backend[74], the basic concepts behind it are described in (Ebner and Palmér, 2008) and (Ebner *et al*., 2009), is used across all Organic.Edunet repositories and stores all information in RDF. This requires that all metadata that are harvested for enriching in the Organic.Edunet repositories are converted from LOM/XML (which is the primary format in most of the source repositories) to an RDF representation. This makes it also possible to freely combine different standards and vocabularies, resulting in enriching LOM metadata with more specific terms from vocabularies such as EUN's LRE and blending in some FOAF and relational predicates from OWL and DC to create interlinkage between resources.

The redesigned Organic.Edunet federation features two different approaches for storing metadata:

1. The distributed repository tools using a triple-store with an abstraction of named graphs and an implementation of the DCMI/IEEE draft using DCAM (Ebner *et al*., 2009), and
2. On the federated portal side, an OWL-based repository based on HP Jena with a relational datastore backend, using an OWL representation of IEEE LOM combined with several ontologies.

The repository tools (based on EntryStore) within the federation might then expose the metadata following the linked data approach according to the aforementioned IEEE/DCMI draft. Indeed, the process of exporting linked data through the portal required the construction of a new module that uses the existing SPARQL endpoint to translate the native RDFOWL representation to the RDF export, combined with a module creating additional RDF links whenever viable in an automated way. Identifiers should be representing at least two types of entities: the object themselves (i.e. the Web contents) and the metadata records. In this case, all the resources are external to the portal and identified by URIs, so it is important to expose only the metadata.

Another attempt to harmonize educational metadata is currently carried out by the Learning Resource Metadata Initiative[75] (LRMI) whose goal is to build a common metadata vocabulary for the description of educational resources. LRMI is led by both the Association of Educational Publishers and the Creative Commons[76]. The applied approach is based on schema.org and has the declared goal of providing mappings to

---

73 The reference implementation is part of EntryStore which is Free Software
74 http://code.google.com/p/entrystore/
75 http://wiki.creativecommons.org/LRMI
76 http://creativecommons.org/

the most common standards for describing education resources, such as LOM and DC.

5.3. Classification and clustering of distributed educational datasets

In this section we survey the methods to perform classification and clustering of distributed data sets and illustrate their advantages both from an end-user perspective (i.e. consuming data) and from an infrastructural perspective (i.e. making the network more interconnected and more accessible).

5.3.1 Clustering

Clustering is a well known machine learning method, used in data mining and knowledge discovery tasks, whose purpose is to classify similar objects into groups (clusters) such that the objects belonging to a cluster are more similar to each other (or closer, in a distance metric space) than they are with objects belonging to other clusters. Typical clustering applications in the context of web data are the discovery of online communities (Lin *et al*., 2010), blogs classification (Yoon *et al.,* 2011), grouping of search results of specialized engines (Vadrevou *et al*., 2011), and segmentation of large collections of documents (Chee and Schatz, 2007). Clustering methods have been applied also to data sets generated from educational settings, to perform tasks such as student modeling, provision of pedagogical support, or understanding resources usage, to name a few (Vellido *et al*., 2010).

Popular clustering methods are the K-means, the fuzzy C-means, and the Kohonen self-organizing map (SOM); in K-means the centroid of each class (cluster) is used to model the data; the fuzzy C-Means assumes that each object in the multidimensional space of the data set can belong to more than one class, with a certain degree; the Kohonen SOM preserves dataset topology, i.e. the distance between classes reflects the distance of their objects in the original multidimensional space. Details about these methods can be found in (Xu and Wunsch II, 2005). In general, clustering performance is affected by the type of distance or similarity metrics adopted and by the nature of the dataset to be clustered.

Traditionally, clustering has been targeted to flat, single-type datasets that can be represented as points in a multi-dimensional vector space; new challenges are posed by the heterogeneous datasets, where relationships between objects are represented through multiple layers of connectivity and similarity (Bolelli *et al*., 2007). These challenges have been addressed by link-based classification methods (Getoor and Diehl, 2005) and by multi-type relational clustering (Li and Anand, 2008); in both contexts the definition of the similarity metrics (or distance) departures from the ones of classic clustering since it takes into account recursively the objects with which each object is related (linked). In link-based classification objects are represented in a data graph consisting of a set of objects connected to each other via a set of links. LinkClus (Yin *et al*., 2006) is a link-based clustering algorithm that exploits the fact, common to many recommender systems, that two items may be deemed similar not only because of pairwise similarity, but because they are linked to similar items. The

performance bottleneck of LinkClus has been ameliorated in a modified version of LinkClus (Yoon, *et al*., 2011). Link-based clustering of an educational data set was first proposed in (Faro and Giordano, 1998) where a web based system of students' design artifacts linked by repurposing information was clustered by the Kohonen SOM, using the links as a means to compute similarity across resources; the clusters enabled the students to use effectively the repository of design projects, thus fostering personal and organizational learning. Various multitype clustering methods (distance based, model based, and spectral clustering) are reviewed in (Li and Anand, 2008), together with techniques for relational object construction and various relational similarity measures; DIVA (Li and Anand, 2007) is a multitype relational clustering framework capable to detect cluster with shapes other than spherical, this flexibility is important since cluster shape is a factor that may affect performance of the clustering algorithm.

These methods assume that the data comes from consistent tables and schemas; when these assumptions do not hold other methods may prove more effective. A recent approach to address heterogeneity of the data set is presented in (Bolelli *et al.*, 2007), where each "block" of multitype information is seen as a source of similarity, these "blocks" taken together, can yield to better clustering results; the method used is K-SVMeans, where K-means is used together with support vector machines (SVM), a supervised classifier that helps preserving relation information when performing the clustering.

Some additional issues must be taken into account when applying clustering to semantic web data, linked data, and, in general, to RDF datasets. Hierarchical clustering (a method that partitions the dataset in clusters organized in a tree) has been applied to ontology-based metadata in (Maedche and Zacharias, 2002), using as similarity measure a combination of taxonomic, relational and attribute values similarities; a result of this study was that relation similarities together with attribute similarities yielded the same clustering as when using only relation similarities. The study in (Grimnes *et al*., 2008) points out that clustering performance is jointly affected by 1) the adopted instance extraction method from the RDF graphs (which expose less structure than the link-bases graphs of objects) and 2) the adopted similarity metric (either based on feature vectors, on graph structure, or on ontologies); the best choice of combination strongly depends on the nature of the data sets. These data sets can be noisy, and can greatly differ based on whether they 1) have been derived from a data base conversion to RDF (typically originating shallow data sets, with only few property relating two resources), 2) are based on a rich ontology, or 3) are generated by crawlings of RDF documents. Findings in (Grimnes *et al.,* 2008) highlight the importance of applying suitable evaluation methods to assess cluster quality, which are very unlikely to be domain-independent, and that conclude that the clustering of RDF resources has still many unanswered questions.

5.3.2 Clustering to support exploratory search of linked educational resources

End-users often engage in a peculiar type of search, known as exploratory search (ES) (Marchionini, 2006). In ES the query is not well focused at the outstart because the users are trying to make sense of an unfamiliar domain whose vocabulary they don't

know; thus exploration of resources is mostly a means to refine the query and to refocus the information needs. Typical approaches to support ES are dynamic taxonomies and facets based interfaces (e.g. Hildebrand *et al*., 2006) that provide a classification (aggregation) of the information space based on explicit properties; this approach has been adopted in the Humboldt Linked Data browser (Kobilarov and Dickinson, 2008) and in a system for exploratory video search (Waitelonis and Sack, 2011).

Clustering can be a different approach to support exploratory search (Giordano *et al*., 2009). The rationale is to create the clusters of related resources based on latent semantic similarities, rather than on explicit properties. Clustering to support the exploratory search of educational resources has been implemented in the mEducator Project (Dietze *et al*., 2012). In particular, clustering is applied to the metadata collected in an RDF store in accordance with the architecture outlined in Figure 1, where data from various repositories are incrementally accrued after having been lifted to a common metadata schema, in this case the mEducator one. The clusters produced by an unsupervised clustering algorithm (either k-means or Kohonen SOM) are used by the application layer to propose to the user, who selects one item from the query results, a set of "related items", i.e. all the items that belong to the same cluster as the selected item. Features are extracted from each metadata instance by processing the free text description fields, taking into account the context of each word (Cohen *et al*., 2010), and then are classified in the similarity space based on a matrix that holds the similarity value between each metadata instance. By operating on the similarity space rather than directly on the feature space, clustering is less sensitive to the specific values of each feature, and is more capable to detect patterns across all features (Faro *et al*., 2009); this property holds also for spectral clustering, a graph theoretic method, especially suitable for multitype data sets, that has been deployed in the educational domain (Trivedi *et al*., 2011). In (Dietze *et al*., 2012) and (Kaldoudi *et al.*, 2011) clustering is provided through a web service that is configurable to allow the selection of the clustering algorithm and the metadata fields to consider for features extraction. This method is similar to the multitype approach of (Bolelli *et al*., 2007), since conceptually the metadata are partitioned in blocks of information that can be selected to contribute to the clustering, and focus the perspective of the clustering (e.g. capturing resources similarities based on the suggested "educational use", rather than on the "subject").

The machine learning clustering approach is orthogonal to the methods that classify resources based on explicit properties (as in faceted search), or on explicit content associations derived by exploiting linked data, as in the exploratory video search of (Waitelonis and Sack, 2011) and in the metadata enrichment of (Dietze *et al*., 2012), where DBpedia is used to find relationships between information instances by mappings terms to LOD and possibly to ontology. In the case of enrichment, clustering is naturally performed on a weighed graph based on the number of enrichment concepts that two nodes have in common.

On the other hand, enrichment could be easily incorporated as an additional source of similarity to take into account for machine learning clustering. The relative benefits of performing clustering either separately on linguistic features and enrichment properties, or in a joint space of linguistic features and enrichments, is an open research question.

Whenever clustering is used as a means to support knowledge discovery, as in the case of ES, issues regarding scalability (Vadrevou *et al*., 2011), cluster quality and interpretability must be considered. Scalability finds a natural solution in the parallelization of the clustering process to exploit Grid or cloud computing technologies, as demonstrated in (Faro *et al*., 2011). Cluster quality evaluation is more difficult, since it requires human input and the properties of the data set should be known a priori; however, quality evaluation is a necessary step to discover experimentally the combination of algorithm, parameters configuration, feature selection methods and similarity metrics, more appropriate to the peculiar properties of the datasets. Interpretability refers to making apparent the reasons why items have been clustered together; methods for this task usually involve re-analysing the class in terms of the original feature space to detect the commonalities; cluster label can be detemined by computing tf-idf (term frequency-inverse document frequency) values to determine the most important terms of each cluster (Bohm *et al*., 2010) or by applying more sophisticated formulae from classic or fuzzy logic.

5.3.3 Clustering as a means to generate new links across repositories

Creating typed links across datasets (interlinking) is a time-consuming process that is receiving increasing attention, as discussed in section 5.2. Although many of the steps involved can be fully automated, automating the whole process can be achieved only for specialized domain, since the task of matching concepts is different depending on the type of data (Woelger *et al*., 2011). In general, the available interlinking tools, surveyed in (Woelger *et al*., 2011), generate as output owl:sameAs triples, and merged datasets; human contribution ranges from specifying the datasets to be linked, to specifying the comparison techniques, to specifying the parameters to be used in the matching methods. Only a few tools go beyond creating "same as" links, and exploit the rdfs:seeAlso property. Among these, Silk (Volz *et al*., 2009) is to date the more flexible one, since the user can provide heuristics to discover whether a given semantic relation exists, e.g. by weighing the similarity metrics to be used. Their approach also defines thresholds for deciding whether a link should be established.

Clustering can be employed as a method, complementary to the above ones, to interlinking repositories and instances of datasets. To understand the advantages of this approach, we analyse two scenarios: 1) an RDF store repository is available, as in Figure 1, where, metadata mapping and lifting has already been performed and there is no need for instance extraction, and 2) no pre-processing of two or more heterogenous, distributed datasets has been performed to achieve schema or instance matching.

In the first case, the information contained in the cluster, as generated in the exploratory search scenario, can be conveniently used to interlink repositories by analyzing the provenance of the items that are clustered together (see Figure 2). Since the semantics is latent, the type of link that can be inferred is of the type rdfs:seeAlso, and this can be declared between instances from the same repository and across repositories. One issue that arises is where this annotation should be kept, and in what cases it should be propagated back to the original data sources. This problem is slightly complicated by the inherent tempo-variant nature of the clustering, according

to which, as long as the number of items to cluster increases or changes, some restructuring of the clusters might take place. Thus a key challenge to make cluster-based interlinking viable is to study suitable cluster metrics to identify cluster's subsets that are likely to remain stable in spite of variations in the dataset. To this end an interesting input can be provided by the methods that facilitate interpretation of the clusters, and by any enrichment information derived from LOD. Also, suitable metrics should be studied to understand when interlinking should be performed not only at the instance level, but also at the repository level, based on the frequency of discovered associations at the instance level.

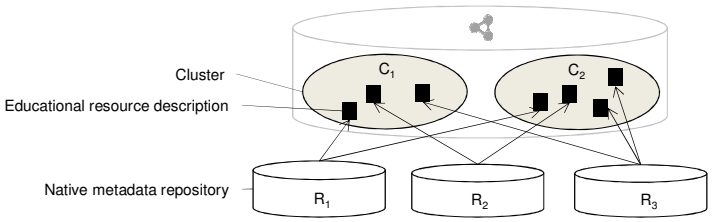

Fig. 2. Clustering as a means to interlink educational resources across independent repositories

This use of clustering shifts the problem of interlinking by determining "overlapping parts" of heterogeneous datasets to the one of interlinking by determining "related parts". Interlinking can be seen as a special case of a link data mining task, i.e., predicting links between two entities, based on their attributes and on other existing links; this is especially challenging because linked datasets are typically sparse, and the prior probability of a link is typically quite small (Getoor and Diehl, 2005). A recent clustering based approach to predict links has employed a variation of the Self-organizing Map (the Probability Measure Graph SOM) (Kc *et al*., 2010) that has the capability to encode cyclic dependencies in a graph; the method has been tested on web documents. If semantically qualified links are being sought, instead of a generic notion of relatedness, some interesting research direction point to the use of supervised learning methods (Lichtenwalter *et al*., 2010).

In the second scenario, where two or more heterogenous, distributed datasets have not been pre-processed to achieve schema or instance matching, clustering offers interesting possibilities. In (Bohm *et al*., 2010) clustering is the initial step of an interactive methodology for profiling LOD, to gather an initial understanding of the data set and overviews of data as a whole. A schema-based similarity measure is used, based on the existence of predicates rather than of values; to each cluster is associated a "mean" schema, which is further refined based on positive and negative association rules that highlight dependencies among predicates. The method in (Partyka *et al*., 2011) uses clustering (in particular, K-medoid clustering, a variant of K-means where the centre of a cluster is an actual data instance) as a way to integrate different sources with diverse schema, possibly without shared instances, into a single unified schema; this can be seen as a way interlinking, with links created at the schema level. In (Hossain *et al*., 2007) a clustering method based on mutual entropy is applied to cluster two distinct datasets that represent the same set of objects, but with different features sets, to obtain information in the clusters that could not be obtained by

clustering each dataset independently. In a similar vein, (Karthikeyani *et al*., 2008) proposes distributed clustering as an emergent method to cluster distributed data sources that might have overlapping features and no shared objects. First local objects at local sites are clustered independently, using fuzzy C-Means, then a unified model is computed and re-applied to the local objects. This approach can be useful when there are reasons (technical, economical, or security) not to transmit the data to a central server, and when data is either horizontally or vertically partitioned. None of these emergent methods have been applied to educational data sets, however they appear to be promising to supplement the metadata mediation and matching illustrated in section 4.3.

5.3.4 Summary

Clustering methods offer a very flexible and powerful tool that, in the context of linked data, can support, directly and indirectly, various educational processes. Educators can be facilitated in discovering resources to design better learning materials or in making sense of data gathered in educational settings; learners have better opportunities to engage in self-directed explorations of information spaces where critical assessment of relevance and quality must be exercised; data sets not meant for educational purposes can be transformed into resources for constructivist learning. Also, from the infrastructural perspective, by trading off some of the precision ensured by semantic matching methods, for a reasonably more generic notion of "relatedness", the interlinking of resources and repositories can be greatly facilitated. However, much work is still to be done, to reach a point where clustering methods can be used as standardized technology. In particular, systematic experimentation is needed to address the complexities of selecting clustering methods, instance extraction methods, and distance/similarity measures best suited to any given domain specific data set. Thus platforms and services that enable this type of experimentation are needed. The mEducator project is a first example in this direction; the development of a metadata schema to track and expose information about clustering as performed on an RDF store of heterogenous resources (Dietze *et al*., 2012) is also a prerequisite to enable more sophisticated forms of interlinking.

6 Conclusion and future work

Integrating existing educational Web resources becomes increasingly important since plenty of data is published openly online with the aim of reuse and Web-scale sharing of resources. While large numbers of competing schemas and interface mechanisms are exploited by individual educational repositories and data collections, we have surveyed and discussed the state of the art in the area. This in particular includes technologies which aim at resolving interoperability problems. Table 1 provides an overview on the surveyed technologies and their categorization into the challenges and principles described in Section 1.

Table 1. Classification of surveyed technologies and their contribution to challenges related to educational Web data integration

	General-purpose	Educational	Relevant Challenges
	Web interfaces, Services & APIs		
Interface mechanisms (APIs & Services) Section 2, 4.1, 4.2	JSON-based feeds, SRU/SRW, REST-ful APIs, OAI-PMH, SQI, SPARQL endpoints, Pubby, WSDL/SOAP-based services	Adoption of general-purpose Web interfaces, OAI-PMH and SQI particularly widely established OKI OSID	*C1 , C2*
Semantic Web Services & services integration Section 2, 3	Minimal Service Model, SAWSDL, WSMO-Lite	"linking hubs" iServe + Smartlink adoption in the mEducator project RDF and DBpedia	*C1 , C2, C3*
	Linked Data		
Standards & Tools Section 3, 4.1, 4.2, 5.2	Linked Data (LD) approach RDF triplestores: Virtuoso, AllegroGraph, HP Jena, Joseki, Sesame, Mulgara, Drupal D2R, Triplify SOAP-based services, SPARQL, use of URIs	LRMI (Learning Resource Metadata Initiative), Luisa (LOM-R ->Ont-Space)	*C1, C2*
Data and Approaches Section 2, 3, 4.1, 5.1, 5.3	Bibliotheque Nationale de France Bioportal, British Museum repository, CNR repository, DBPedia, EdNA Online, Europeana project, Linked Life Data, PubMed, VOAR project	Ariadne RDF, eduSource Canada project EduLearn project, LACLO, Linked Education, KOCW, GLOBE, LORNET, LOP2P plugin, LODUM project (Linked Open Data at the University of Muenster), LRE, mEducator project, OER, OERCommons, OpenLearn (OU Linked Data store), Organic.Edunet project, Oxford University platform, OUJ, Southampton University repository	*C1*
Linked Data integration/interlinking/federation Section 5.2	Apache Stanbol framework DBPedia Spotlight OpenCalais	EntryStore backend makes it possible to freely combine different standards and vocabularies, resulting in enriching LOM metadata with more specific terms from vocabularies such as EUN's LRE and blending in some FOAF and relational predicates from OWL and DC to create interlinkage between resources.	*C3, C4*
	Open Educational Resources – Standards and Approaches		
Repositories Section 4.1, 4.2		ARIADNE, Connexions, Edna (APIs) EduSource Communications Layer (ECL) IMS DRI model, MERLOT, Merlot (Web services), MITOpenCourseWare, OCW Search Meta Data (API), OpenCourseFinder, OpenCourseWare Search (API)	*C3*
Metadata schemas Section 5.1, 5.2	Dublin core (DC) DCAM, DCMI/IEEE	ADL SCORM, CanCore, DC-Ed, GEM, IEEE Learning Object Metadata (LOM), IMS, ISO/IEC MLR, UK LOM Core.	*C1, C3*
	Mapping, clustering and interlinking		
Schema mapping Section 4.2, 4.3,	COMA++, COMA 3.0, S-Match, RiMOM,	Application of SmartLink & iServe in mEducator: using lifting-templates for schema mapping	*C3*
Clustering and interlinking Section 5.3	Link-based clustering: LinkClus, modified LinkClus, DIVA, spectral clustering, Kohonen SOM, K-SVMeans RDF clustering: K-means, Kohonen SOM, hierarchical clustering Interlinking: K-means, PMGraph SOM, supervised learning Dataset integration: K-means, K-medoid clustering; Fuzzy C-means	Application of clustering techniques within educational contexts (eg mEducator)	*C3, C4*

While there is a wide variety of technologies available dealing with exposing, sharing and integrating educational Web data, it can be noted that specifically, more recentand Linked Data-based approaches have gained a lot of momentum and started

realising the vision of highly accessible and Web-wide reusable OER by providing the standards, tools, and Web infrastructure to expose and interlink educational data at Web-scale. That was demonstrated and proven by the vast amounts of educational metadata collections and university data which have been provided throughout the last years according to Linked Data principles (see, for instance, Sections 5.1and 5.2).

In addition, the Linked Data approach has provided a vast body of knowledge which, though not of explicit educational nature, offers significant potential for exploitation in educational contexts. This includes cross-domain datasets (such as DBpedia) as well as domain-specific vocabularies which provide formal descriptions of domain knowledge or domain-specific data collections (such as PubMed or Europeana).

To this end, the Linked Data approach had and will have a strong impact on the educational field and has already started to replace the fragmented landscape of educational technologies and standards with a more unified approach, which allows to integrate and interlink educational data of any kind. Here, one of the particular strengths of the Linked Data approach is the fact that Linked Data does not impose common and shared schemas but instead, accepts heterogeneity and offers solutions by fundamentally relying on links between disparate schemas and datasets to facilitate Web-scale interoperability.

References

Amuller, D., Do, H.-H., Massmann, S. and Rahm, E. (2005), "Schema and Ontology Matching with COMA++", *Proceedings of the 24th International Conference on Management of Data / Principles of Database Systems (SIGMOD),* June 13-16, 2005, Baltimore, Maryland, USA.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007), "DBpedia: A nucleus for a web of open data", *The Semantic Web Journal.*

Baldoni, M., Baroglio, C., Brunkhorst, I., Henze, N., Marengo, E. and Patti, V. (2006), "A Personalization Service for Curriculum Planning", *Proceedings of the 14th Workshop on Adaptivity and User Modeling in Interactive Systems*, Hildesheim.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001), "The Semantic Web", *Scientific American Magazine.* May 2001, Vol. 284 No. 5, pp. 34-43.

Bernstein, P.A., Madhavan, J. and Rahm, E. (2011), "Generic Schema Matching, Ten Years Later", *PVLDB Endowment*, 2011. Vol. *4* No. 11, pp. 695-701.

Bizer, C., Heath, T. and Bernes-Lee, T. (2009), "Linked data - The Story So Far. Special Issue on Linked data", *International Journal on Semantic Web and Information Systems (IJSWIS).*

Bizer, C., Heath, T., Idehen, K. and Berners-Lee, T. (2008), "Linked data on the web (LDOW2008)", *Proceedings of the 17th international conference on World Wide Web* (WWW '08), April 21-25, 2008, Beijing, China.

Bohm, C., Naumann, F., Abedjan, Z., Fenz, D., Grutze, T., Hefenbrock, D., Pohl, M. and Fenz, D. (2010), "Profilink Linked Open data with ProLOD", *Proceedings of the IEEE ICDE Workshops*, March 1-6, 2010, pp. 175-8.

Bolelli, L., Ertekin, S., Zhou, D. and Giles, C.L. (2007), "K-SVMeans: A hybrid clustering algorithm for multitype interrelated datasets", *Proceedings of the IEEE WIC/ACm Conf. on Web Intelligence,* November 2-5, 2007, Silicon Valley, California, USA, pp. 198-204.

Carvalho, M. G., Laender, A. H., Gonçalves, M. A. and da Silva, A. S. (2008), "Replica identification using genetic programming", *Proceedings of the ACM SAC (Symposium on Applied Computing),* March 16-20, 2008, Ceará, Brazil, pp. 1801-6.

CEN (2005), "A simple query interface specification for learning repositories". *CEN Workshop Agreement (CWA 15454),* available at: ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/WS-LT/cwa15454-00-2005-Nov.pdf (accessed October 6, 2011).

Chee, B. and Schatz, B. (2007), "Document clustering using small world communities", *Proceedings of the Joint Conference on Digital Libraries (JCDL'07),* June 18-23, 2007, Vancouver, British Columbia, Canada, ACM, pp. 53-62.

Cheung, K.-H., Frost, H. R., Marshall, M. S., Prud'Hommeaux, E., Samwald, M., Zhao, J. and Paschke, A. (2009), "A journey to Semantic Web query federation in the life sciences", *BMC bioinformatics,* Vol. 10 (Suppl 1):S10.

Cohen, T., Schvanevendt, R. and Widdows, D. (2010), "Reflective random indexing and and indirect inference: a scalable method for discovery of implicit connections", *J. Biomedical Informatics*, Vol. 43 No. 2, pp. 240-56.

Dhamankar, R., Lee, Y., Doan, A., Halevy, A. and Domingos, P. (2004), "iMAP: discovering complex semantic matches between database schemas", *Proceedings of the ACM SIGMOD international conference on Management of data (SIGMOD '04),* New York, USA, pp. 383-94.

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

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

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

Do, H.-H. and Rahm, E. (2002), "COMA - A System for Flexible Combination of Schema Matching Approaches", paper presented at the *28th VLDB Conference,* Hong Kong, China, 2002.

Doan, A. and Halevy, A. (2005), "Semantic Integration Research in the Database community: A Brief Survey", *AI magazine*, Vol. 26, pp.83–94, 2005.

Downes, S. (2004), "EduSource: Canada's Learning Object Repository Network", *International Journal of Instructional Technology and Distance Learning*, Vol. 1 No. 3, available at: http://www.itdl.org/Journal/Mar_04/article01.htm

Duval, E., Hodgins, W., Sutton, S. and Weibel, S. (2002), "Metadata Principles and Practicalities", *D-Lib Magazine*, Vol. 8 No. 4, doi:10.1045/april2002-weibel.

Ebner, H., Manouselis, M., Palmér, M., Enoksson, F., Palavitsinis, N., Kastrantas, K. and Naeve, A. (2009), "Learning Object Annotation for Agricultural Learning Repositories", *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Riga, Latvia, 2009.

Ebner, H. and Palmér, M. (2008), "A Mashup-friendly Resource and Metadata Management Framework", in Wild, Kalz, and Palmér (Eds.), *Mash-Up Personal Learning Environments,* Proceedings of the 1st Workshop MUPPLE, European Conference on Technology Enhanced Learning (EC-TEL), Maastricht, The Netherlands, CEUR Vol. 388, available at: http://ceur-ws.org/Vol-388/

Faro, A. and Giordano, D. (1998), "StoryNet: an evolving network of cases to learn information systems design", *IEEE Software*, Vol. 145 No. 4, pp. 119-27.

Faro, A., Giordano, D., Maiorana, F. and Spampinato, C. (2009), "Discovering genes-diseases associations from specialized literature using the grid", *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13 No. 4, pp. 554-60.

Faro, A, Giordano, D. and Maiorana, M. (2011), "Mining massive datasets by an unsupervised parallel clustering on a GRID: Novel algorithms and case study", *Future Generation Computer Systems*, Vol. 27 No. 6, pp. 711-24.

Getoor, L. and Diehl, C.P. (2005), "Link mining: a survey", *SIGKDD Explorations*, Vol. 7 No. 2, pp.3-12.

Giordano, D. Faro, A., Maiorana, F., Pino, C. and Spampinato, C. (2009), "Feeding back learning resources repurposing patterns into the "information loop": opportunities and challenges", *Proceedings of the 9th Int. Conf. on Information Technology and Applications in Biomedicine (ITAB*), November 5-7, 2009, Larnaca, Cyprus.

Giunchiglia, F., Autayeu, A. and Pane, J. (2010), "S-Match: an open source framework for matching lightweight ontologies", *Semantic Web Journal, 2010.*

Grimnes, G. A., Edwards, P. and Preece, A. (2008), "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. LNCS 502, Springer-Verlag. pp. 303-17.

Haslhofer, B. and Schandl, B. (2008), "The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data", *Proceedings of the International Workshop on Linked Data on the Web (LDOW2008)*, co-located with WWW 2008, April 22, 2008, Beijing, China.

Hatala, M., Richards, G., Eap, T. and Willms, J. (2004), "The interoperability of learning object repositories and services: standards, implementations and lessons learned", *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters (WWW Alt. '04*), ACM, New York, NY, USA, pp. 19-27.

Henze, N. (2006), "Personalized E-Learning in the Semantic Web". Extended version of 4, *International Journal of Emerging Technologies in Learning (iJET)*, Vol. 1 No. 1.

Henze, N., Dolog, P. and Nejdl, W. (2004), "Reasoning and Ontologies for Personalized E-Learning", *Educational Technology & Society*, Vol. 7 No. 4, pp. 82-97.

Hildebrand, M., van Ossenbruggen, J. and Hardman, L. (2006), "/facet: a browser for heterogeneous semantic web repositories", *Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, LNCS 4273, Springer, pp. 272-85.

Hilera, J.R., Oton, S., Ortiz, A., de Marcos, L., Martinez, J.J., Gutierrez, J.A., Gutierrez, J.M. and Barchino, R. (2009), "Evaluating simple query interface compliance in public repositories", *Proceedings of the 9th IEEE International Conference on Advanced Learning Technologies,* July 15-17, 2009, Riga, Latvia.

Hossain, M., Bridges, S., Wang, Y. and Hodges, J. (2007), "Extracting partitional clusters from heterogenous datasets using mutual entropy", *Proceedings of the IEEE Int. Conf. on Information Reuse and Integration (IRI),* August 13-15, 2007, Las Vegas, Nevada, USA, pp. 447-54.

IEEE (2002), "IEEE Standard for Learning Object Metadata", *IEEE Std 1484.12.1-2002*, pp.i–32. doi: 10.1109/IEEESTD.2002.94128.

IMS DRI (2003), "IMS Digital Repositories Core Functions Information Model", Version 1, available at: www.imsglobal.org/digitalrepositories/ (accessed 20 October 2011).

Kaldoudi, E., Dovrolis, N., Giordano, D. and Dietze, S. (2011), "Educational Resources as Social Objects in Semantic Social Networks", *Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age (Linked Learning 2011*) at the 8th Extended Semantic Web Conference (ESWC), May 29, 2011, Heraklion, Crete.

Kalfoglou, Y. and Schorlemmer, M. (2003), "Ontology mapping: the state of the art", *Knowl. Eng. Rev*. 18, 1 (January 2003), pp. 1-31.

Karthikeyani Visalkshi, N., Thangavel, K. and Alagambigai, P. (2008), "Distributed clustering for data sources with diverse schema", *Proceedings of the 3rd Intern Conference on Convergence and hybrid information technolog,*. IEEE Computer Society Washington, DC, USA, pp. 1058-63.

Kc, M., Chau, R. and Hagenbuchner, M. (2010), "A machine learning approach to link prediction for interlinked documents", *Proceedings of the INEX 2009 Workshop*, LNCS 6203, pp. 342-54.

Klemke, R., Ternier, S., Kalz, M. and Specht, M. (2010), "Implementing infrastructures for managing learning objects", *British Journal of Educational Technology*, Vol. 41, pp. 873-82.

Kobilarov, G. and Dickinson, I. (2008), "Humboldt: Exploring Linked Data", *Proceedings of the Linked Data on the Web (LDW'08) Workshop,* April 2008, Beijing, China.

Kopecky, J., Vitvar, T. and Gomadam, K. (2008), "MicroWSMO", Deliverable, *Conceptual Models for Services Working Group*, available at: http://cms-wg.sti2.org/TR/d12/v0.1/20090310/d12v01_20090310.pdf.

Koutsomitropoulos, D.A., Alexopoulos, A.D., Solomou, G.D. and Papatheodorou, T.S. (2010), "The Use of Metadata for Educational Resources in Digital Repositories:

Practices and Perspectives", *D-Lib Magazine*. January/February 2010, Vol. 16 No. 1/2, available at: http://www.dlib.org/dlib/january10/kout/01kout.print.html#14

Lagoze, C. and Van de Sompel, H. (2001), "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework", *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01),* June 24-28, 2001, Roanoke, VA, USA.

Lagoze, C. and Van de Sompel, H. (2003), "The making of the Open Archives Initiative Protocol for Metadata Harvesting", *Library Hi Tech*, Vol. 21 Iss: 2, pp.118-28.

Li, T. and Anand, S.S. (2008), "Multi-type Relational Clustering Approaches: Current State-of-the-Art and New Directions", *Proceedings of the International Conference on Intelligent Systems and Networks (IISN 2008)*, February 22-24, 2008, Klawad, Jagadhari, India.

Li, T. and Anand, S.S. (2007), "DIVA: A Variance-based Clustering Approach for Multi-type Relational Data", *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07),* November 6-9, 2007, Lisbon, Portugal.

Li, J., Tang, J., Li, Y. and Luo, Q. (2009), "RiMOM: A dynamic multistrategy ontology alignment framework", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 21, pp. 1218-32.

Lichtenwalter, R.N., Lussier, J.T. and Chawla, N.V. (2010), "New perspectives and methods in link prediction", *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10),* July 25-28, 2010, Washington DC, DC, USA.

Lin, C., Koh, J. and Chen, A. L.P. (2010), "A better strategy of discovering link pattern based communities by classical clustering methods", *Proceedings of the 14th Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD 2010)*, June 21-24, 2010, Hyderabad, India, Springer, Part I, LNAI 6118, pp.56-67.

Maedche, A. and Zacharias, V. (2002), "Clustering ontology-based metadata in the semantic Web", *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'02)*, Springer-Verlag, pp. 348-60.

Marchionini, G. (2006), "Exploratory search: from finding to understanding", *Communication ACM.* New York, NY, USA, Vol. 49 No. 4, pp. 41-46.

Massmann, S., Raunich, S., Aumuller, D., Arnold, P. and Rahm, E. (2011), "Evolution of the COMA match system", *Proceedings of the 6th International Workshop on Ontology Matching (OM-2011)*, CEUR-WS, Vol. 814, 2011.

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

Morgan, E. L. (2004), "An Introduction to the Search/Retrieve URL Service (SRU)", *Ariadne 40*, available at: www.ariadne.ac.uk/issue40/morgan/ (accessed October 10, 2011).

Nilsson, M. (2010), *From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization*. PhD thesis, KTH Royal Institute of Technology, Sweden, 2010.

Nunes, B. P., Caraballo, A., Casanova, M. A., Breitman, K. K. and Leme, L. A. P. P. (2011), "Complex matching of RDF datatype properties", *Proceedings of the 6th International Workshop on Ontology Matching (OM-2011)*, CEUR-WS, Vol. 814, 2011.

Ochoa, X. and Duval, E. (2009), "Quantitative analysis of learning object repositories", *IEEE Transactions on Learning Technologies*, Vol. 2 No. 3, pp. 226-38.

Partyka J., Khan, L. and Thuraisingham, B. (2011), "Semantic schema matching without shared instances", *Proceedings of the 5th Int. Conference on semantic computing*. September 18-21, 2011, Palo Alto, CA, USA, pp. 297-302.

Prakash, L. S., Saini, D. K. and Kutti. N. S. (2009), "Integrating *EduLearn* learning content management system (LCMS) with cooperating learning object repositories (LORs) in a peer to peer (P2P) architectural framework", *SIGSOFT Softw. Eng. Notes.* Vol. 34 No. 3 (May 2009), pp. 1-7. DOI=10.1145/1527202.1527212

Rahm, E. and Bernstein, P. (2001), "A survey of approaches to automatic schema matching", *The VLDB Journal*, Vol. 10 No. 4, pp. 334-50.

Raunich, S. and Rahm, E. (2011), "ATOM: Automatic target-driven ontology merging", *Proceedings of the 27th International Conference on Data Engineering (ICDE '11)*, IEEE Computer Society, Washington, DC, USA, pp. 1276-79.

De Santiago, R. and Raabe, A.L.A. (2010), "Architecture for Learning Objects Sharing among Learning Institutions-LOP2P", *IEEE Transactions on Learning Technologies*, April-June, 2010, pp. 91-5.

Schmidt, A. and Winterhalter, C. (2004), "User Context Aware Delivery of E-Learning Material: Approach and Architecture", *Journal of Universal Computer Science (JUCS),* Vol. 10 No. 1.

Sheth, A. P., Gomadam, K. and Ranabahu, A. (2008), "Semantics enhanced services: Meteor-s, SAWSDL and SA-REST". *IEEE Data Eng. Bul* l., Vol. 31 No. 3, pp. 8-12.

Shvaiko, P. and Euzanat, J. (2005), "A survey of schema-based matching approaches", *Journal on Data Semantics,* IV, pp. 146-71, 2005.

Sicilia, M.A, Sanchez-Alonso, S., Alvarez, F., Abián, A. and Garcia-Barriocanal, E. (2011), "Navigating learning Resources through Linked Data: a preliminary Report on the Re-Design of Organic.Edunet", in *Linked Learning 2011,* Proceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, at the 8th Extended Semantic Web Conference, ESWC2011. May 29, 2011, Heraklion, Greece, CEUR-WS, Vol. 717, available at: http://ceur-ws.org/Vol-717/

Schmidt, A. (2005), "Bridging the Gap Between E-Learning and Knowledge Management with Context-Aware Corporate Learning (Extended Version)", *Proceedings of the Professional Knowledge Management conference (WM 2005)*, April 10-13, 2005, Kaiserslautern, Germany.

Simon, B., Dolog, P., Miklós, Z., Olmedilla, D. and Sintek, M. (2004), "Conceptualising Smart Spaces for Learning", *Journal of Interactive Media in Education*. 2004(9), available at:http://www-jime.open.ac.uk/2004/9.

Simon, B., Massart, D., Assche, F., Ternier, S., Duval, E., Brantner, S., Olmedilla, D. and Miklos. Z. (2005), "A simple query interface for interoperable learning repositories", *Proceedings of the 1st Workshop On Interoperability of Web-Based Educational Systems*, May 10-14, 2005, Chiba, Japan.

Ternier, S., Bosman, B., Duval, E., Metzger, L., Halm, M., Thorne, S. and Kahn, J. (2006), "Connecting OKI And SQI: One Small Piece Of Code, A Giant Leap For Reusing Learning Objects", in E. Pearson & P. Bohman (Eds), *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006,* Chesapeake, VA: AACE, pp. 825-31.

Ternier, S., Massart, D., Campi, A., Guinea, S., Ceri, S. and Duval, E. (2008), "Interoperability for Searching Learning Object Repositories: The ProLearn Query Language", *D-Lib Magazine*, Vol. 14 No. 1/2.

Trivedi, S., Pardos, Z. A., Sarkozy, G. N. and Heffernan, N. T. (2011), "Spectral clustering in educational data mining", *Proceedings of the 4th International Conference on Educational Data Mining,* July 6-8, 2011, Eindhoven, the Netherlands.

Vadrevou, S., Teo, C.H., Rajan, S., Punera, k., Dom, B., Smola, A., Chang, Y. and Zheng,, Z. (2011), "Scalable clustering on news search results", *Proceedings of the WSDM'11*, February 09-12, 2011, Kowloon, Hong Kong, ACM, pp. 675-83.

Vellido, A., Castro, F. and Nebot, A. (2010), *Clustering educational data, In Handbook of Educational Data Mining*, CRC Press, pp. 75-92.

Vitvar, T., Kopecky, J., Viskova, J. and Fensel, D. (2008), "WSMO-lite annotations for web services", in Hauswirth, M., Koubarakis, M., and Bechhofer, S., (Eds), *Proceedings of the 5th European SemanticWeb Conference, LNCS*. Berlin, Heidelberg: Springer Verlag.

Volz, J., Bizer, C., Gaedke, M. and Kobilarov, G. (2009), "Silk - a link discovery framework for the web of data", *Proceedings of the Linked Data on the Web Workshop (LDOW2009),* April 20, 2009, Madrid, Spain, CEUR-WS, Vol. 538, available at: http:// ceur-ws.org/Vol-538/ldow2009_paper13.pdf.

Waitelonis, J. and Sack, H. (2011), "Towards exploratory video search using linked data", *Multimedia Tools and Applications*, Vol. 53 (2011), pp. 1-28.

Woelger, S., Siorpaes, K., Buerger, T., Simperl, E., Thaler, S. and Hofer, C. (2011), *A Survey on Data Interlinking Methods*, Technical report, Semantic Technology Institute, March 2011.

World Wide Web Consortium (2008). *W3C Recommendation, SPARQL query language for RDF*, 2008, available at:www.w3.org/TR/rdf-sparql-query/

Xu, R. and Wunsch II, D. C. (2005), "Survey of clustering algorithms", *IEEE Transactions on Neural Networks,* Vol. 16 No. 3, pp. 645-78.

Yin, X., Han, J. and Yu, P.S. (2006), "LinkClus: efficient clustering via heterogeneous semantic links", *Proceedings of the 32nd international conference on Very large data bases (VLDB '06),* September 12-15, 2006, Seoul, Korea, pp. 427-38.

Yoon, S., Song, S. and Kim, S. (2011), "Efficient link-based clustering in a large scaled blog network", *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*. February 21-23, *2011*, Seoul, Korea, ACM, pp. 1-5.

Yu, H. Q., Dietze, S., Li, N., Pedrinaci, C., Taibi, D., Dovrolls, N., Stefanut, T., Kaldoudi, E. and Domingue, J. (2011), "A linked data-driven & service-oriented architecture for sharing educational resources", in *Linked Learning 2011,* Proceedings of the 1st International Workshop on eLearning Approaches for Linked Data Age, at the 8th Extended Semantic Web Conference (ESWC2011), May 29, 2011, Heraklion, Greece.

Zablith, F., d'Aquin, M., Brown, S. and Green-Hughes L. (2011b), "Consuming Linked Data Within a Large Educational Organization", *Proceedings of the 2nd International Workshop on Consuming Linked Data (COLD)* at International Semantic Web Conference (ISWC), October 23-27, 2011. Bonn, Germany.

Zablith, F., Fernandez, M. and Rowe, M. (2011a), "The OU Linked Open Data: Production and Consumption", in *Linked Learning 2011, P*roceedings of the 1st International Workshop on eLearning Approaches for the Linked Data Age, at the 8th Extended Semantic Web Conference (ESWC), May 29, 2011, Heraklion, Crete.