SemUNIT - French UNT and Linked Data

Yoann Isaac, Yolaine Bourda, Monique Grandbastien

SemUNIT – French UNT and Linked Data

Yoann Isaac, Yolaine Bourda
Supélec
Plateau de Moulon
91192 Gif-sur-Yvette Cedex, France
{Yoann.Isaac,
Yolaine.Bourda}@supelec.r

Monique Grandbastien
université de Lorraine
bat. LORIA, campus scientifique,
BP 239, 54506 VANDOEUVRE Cedex, France
Monique.Grandbastien@loria.fr

ABSTRACT

Over the past 15 years, the explosion in the number of learning materials available on the Web has raised the problem of their sharing. For several years, learning resources are annotated with metadata to ease this sharing. On the other hand, Semantic Web and Linked Data approach provide tools to publish metadata in a standardized way, allowing data to be shared and reused across applications, enterprises, and community boundaries.

In this paper, we describe the SemUnit project, initiated by french higher education institutions. This project aims at taking advantages of Semantic Web and Linked Data to improve e-learning services for a wide set of french higher education institutions. We present, firstly, the ontology designed to support the project: an OWL ontology taking into account the semantics of LOM elements. Afterwards, we present our architecture and some semantic services[1].

Categories and Subject Descriptors

I.2.4 [**ARTIFICIAL INTELLIGENCE**]: Knowledge Representation Formalisms and Methods; K.3.1 [**COMPUTERS AND EDUCATION**]: Computer Uses in Education

General Terms

Design

Keywords

E-learning, ontology for learning resources, Semantic Web, Linked Data

1. INTRODUCTION

Today, more and more learning materials are available on the Web. The sharing of these learning materials can greatly ease the creation of quality courses, especially when they are composed of well-defined reusable learning objects [9]. This sharing requires the creation of tools allowing to store, search and evaluate them efficiently. The full text search approach mostly used today is not adapted. Consequently, for several years, learning materials have been annotated with metadata using mainly the Learning Object Metadata (LOM) standard, one of the LOM application profiles or the Dublin Core (DC).

The creation of high quality learning objects is time consuming and requires to be supported. In 2003, the French government created the UNTs (Universités Numériques Thématiques: Thematic Digital Universities) in order to support the creation of high quality learning material and ease their sharing between higher education institutions (Universities and "Grandes Ecoles"). Each UNT gathers learning resources belonging to a specific field. Among these UNT, UNIT[2] is the one dedicated to "Engineering sciences and technologies". UNIT currently provides a free access to several thousands of high quality learning objects. They are annotated with metadata complying to the SupLOMFR schema[3], a LOM application profile dedicated to French higher education institutions. Metadata are published thanks to a network of OAI-PMH[4] portals under the open-source software ORI-OAI[5]. This system makes metadata and resources harvestable [13] and then allows the creation of cross-repository services. Nonetheless, only search services based on input forms are available. Although useful, this kind of service is quite restrictive and does not allow indexing of educational material by search engines. This therefore restricts the visibility of the metadata and, of the described resources and thus reduces their use (to be used a resource first needs to be found and known). Furthermore, it is difficult to automatically link those metadata with results coming from other repositories (like DBLP for example).

On the other hand, the explosion of the Semantic Web and Linked Data, in the last few years, has allowed the production of a large number of data with a clear semantic, increasing the visibility of such data. The expression "Linked Data" corresponds to a set of best practices for publishing structured data [6]. In the e-learning context, the respect of linked data principles can be beneficial in many ways. Indeed, if learning resources belonging to several repositories are linked together or with others resources: cross-repository services can be created (like search engines); learning re-

[1]http://semunt.supelec.fr/portal/

[2]http://www.unit.eu/fr

[3]http://www.sup.lomfr.fr/

[4]http://www.openarchives.org/OAI/openarchivesprotocol.html

[5]http://www.ori-oai.org/display/ORIOAI/ORI-OAI.ORG

Linked Learning 2012, 17 April 2012, Lyon, France.

sources can reference RDF resources close to theirs topics to offer useful information (DBpedia [14]); learning resources can be described by linking them to specific taxonomies; learning ontologies can be mapped to others and these mapping can be used efficiently for reasoning.

Thus, the SemUnit project has been launched in order to bring the advantages of Semantic Web and Linked Data technologies to the French UNT repositories. This relies on the creation of an ontology taking into account the semantics of LOM elements. But unlike other works, the designed ontology has a structure (classes and properties) that is not based on the LOM table and reflects the meaning of metadata elements.

The main steps of the SemUnit project are: the creation of an ontology capturing the semantics of the metadata, the automatic translation of metadata from XML to RDF triples, the study of services taking advantage of the Semantic Web metadata previously created, the definition of an architecture supporting the desired services and the creation or use of the related tools.

This paper is organised as follow: section 2 describes the SemUNT ontology, section 3 presents the architecture and the tools, section 4 introduces some services grouped in a Web application, section 5 presents some related works and finally, section 6 concludes and presents some future works

2. THE SEMUNT ONTOLOGY

UNT learning materials are described by metadata complying with the SupLOMFR schema (an IEEE LOM application profile) and encoded in XML. Give meanings to these annotations and transform them to RDF triples requires the creation of an ontology. OWL rather than RDF(S) has been chosen to implement this ontology because of it's greater expressiveness, for example OWL cardinality restrictions are particularly important to express cardinalities. Furthermore, unlike other works, the designed ontology has a structure (classes and properties) that is not based on the LOM table and reflects the meaning of the metadata elements.

The next two subsections describe the choices made to create this ontology. The first part deals with the general structure of the ontology and the second with the management of vocabularies.

2.1 SemUNT ontology structure

2.1.1 Related ontologies

Several e-learning ontologies based on the LOM or RDF(S) LOM bindings have already been created. The first was defined in RDF(S) by Nilsson [10]. It proposes some interesting ideas (which have been adopted for the SemUNT ontology): use language tags for lom:langString, represent VCARD text with a specific ontology, integrate Dublin Core Properties (title, format...) and DataType (DateTime, Language...) when possible, etc. Nevertheless, it does not provide solutions concerning classifications or vocabularies management and its RDF(S) implementation does not allow cardinality restrictions.

More recently, other LOM ontologies have been created (e.g. [2, 3, 4, 8]). Some of these ontologies ([3, 4]) are very closed to the LOM's table structure and then, allow an easy transformation from a LOM XML binding to an RDF one. The semantics of the LOM elements is, on the contrary, not well modelled and some ontology's classes have

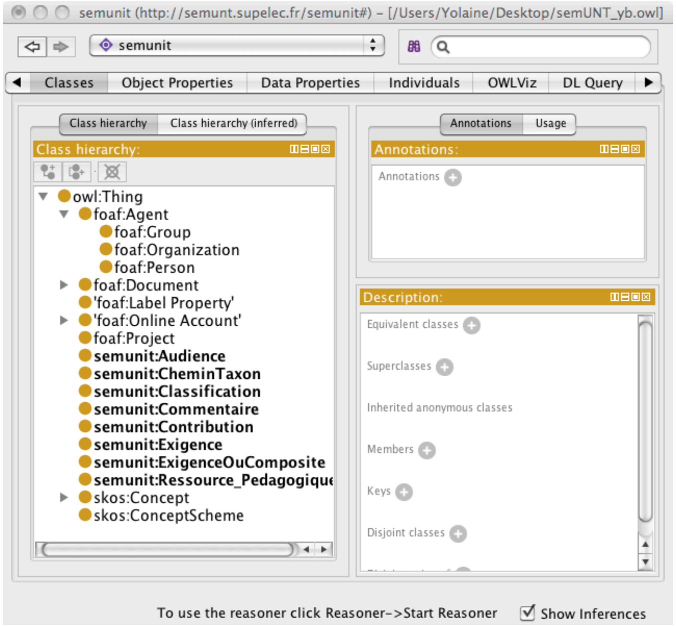

Figure 1: Classes of the SemUNT ontology

no real meaning. LOM categories are for instance mapped to classes even if this mapping is artificial. Only the relation category is mapped differently (e.g. in [4]) to respect its meaning: LOM relations allow to link learning resources to others and consequently are mapped to object properties. Other ontologies [2, 8] only consider a subset of the LOM and although their choices are closer to ours we have not been able to reuse them as is because we needed the entire LOM and because we wanted to use OWL (to express constraints) and not RDF(S) and SKOS to express controlled vocabularies.

2.1.2 Ontology Construction

The SemUNT ontology (see figures 1 and 2) has been created with some objectives and constraints in mind: respect the meaning of metadata elements; maximize its interoperability; respect linked data rules. This last point led us to avoid some RDF constructions like collections, containers and blank nodes.

Classes.

All learning resources are instances of the class LearningResource. Classes have been defined for categories representing something real for learning resources like Classification. Classes have also been defined for some structured elements like Contribution because a contribution is a grouping of several attributes: a date, a role and a contributor.

Properties.

Simple elements are mapped to properties: datatype properties for these (title, description, format, etc) taking their values in data types and object properties for those taking their values in vocabularies (see section 2.2). Furthermore, for each of the defined classes (Classification, Contribution...) an object property has been created (hasContribution, hasClassification...) whose domain is the class LearningResource and range is the defined class.

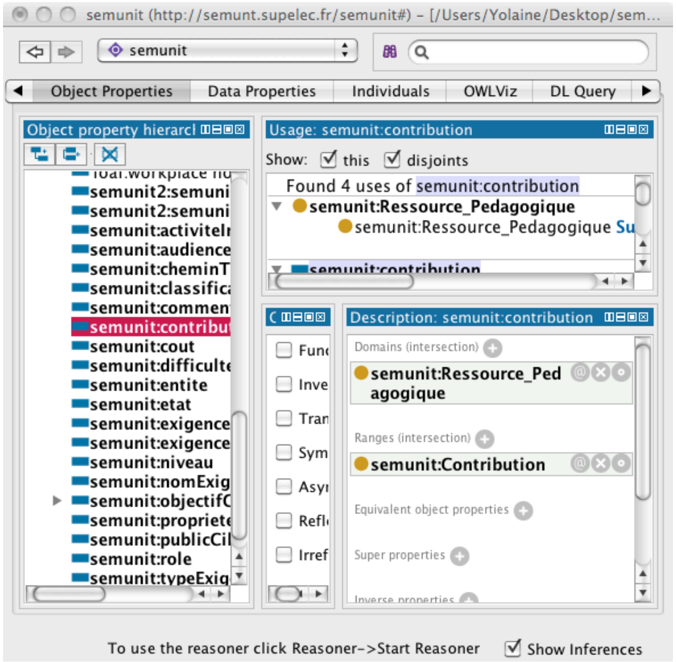

Figure 2: Properties of the SemUNT ontology

The category Relation is mapped to an object property linking two learning resources. The kind of relation is defined using the subProperty construct. This representation allows a better browsing of learning resources and stays close to the meaning of the metadata element. However, with this representation, the description of the relation is lost but this element is rarely filled in with UNT learning resources. This solution has been nevertheless chosen because the pros outweigh the cons.

Interoperability.

The SupLOMFR schema, like the LOM one, defines some correspondences with Dublin Core (DC) properties. We choose to integrate them directly in the SemUNT ontology (identifier, title , language, coverage, type, rights, relation, description, format and subject).

The ontology interoperability is also improved by using the FOAF Ontology, one of the most used for people's description [1]. Thus the foaf:Agent class is used for VCard elements. The only mapping problem is that some VCard attributes does not correspond to any FOAF properties for now but these elements are not filled in UNT metadata.

2.2 Vocabularies management

Some SupLOMFR metadata elements take their value in controlled vocabularies. Some of the possible values are coming from the LOM Schema, some of them are new values linked or not to some LOM values by a relationship (like narrower). For example in the LOM, for the 5.6 level element we have the value "Higher Education", in the SupLOMFR we have "Enseignement Supérieur" (the same than the LOM one but in French), and more specific values like "Licence" or "Doctorat". But as we also want to be able to use metadata complying with the LOM schema (and other LOM application profiles like normetic), we decided to use SKOS (Simple Knowledge Organization System) to describe vocabularies. In this way, one can define distinct controlled vocabularies sharing concepts or having concepts linked by relationships (like broader or narrower). An example of the implementation of the vocabularies is given figure 3.

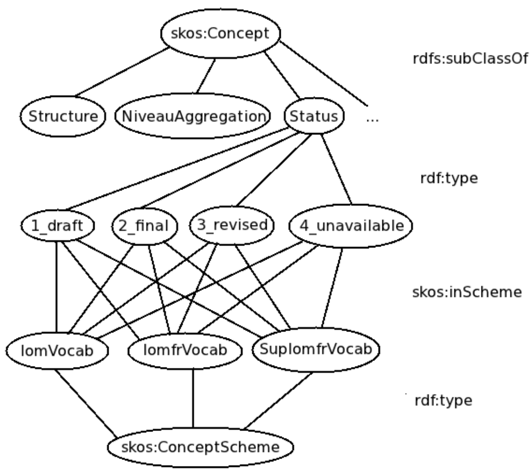

Figure 3: Vocabularies in SKOS

3. ARCHITECTURE AND TRANSFORMATION TOOLS

3.1 Architecture

Concerning the storage of RDF triples, the system requirements were the following ones: it should accept at least 1 millions RDF triples and be scalable; it should offer a SPARQL access with relatively fast answer; it should propose a linked data interface;

To meet these requirements, the following tools have been chosen: OWLIM-Lite[6] for storing RDF triples, Sesame2[7] for managing this repository and Pubby[8] for exposing a linked data interface. OWLIM was choosen because of it's good results at the Berlin SPARQL Benchmark[9]. The architecture is presented in figure 4.

3.2 Transformation tools

We defined transformation tools allowing to automatically convert XML metadata to RDF triples (complying with the ontology previously defined) and studied RDF storage solutions (1 million RDF triples have been created from 6456 UNT metadata files).

An handmade XSLT transformation has been chosen to automatically "RDFize" the XML UNT metadata although different "RDFizers" are already freely available. Indeed,

[6]http://www.ontotext.com/owlim
[7]http://www.openrdf.org/
[8]http://www4.wiwiss.fu-berlin.de/pubby/
[9]http://www4.wiwiss.fu-berlin.de/bizer/berlinsparqlbenchmark/

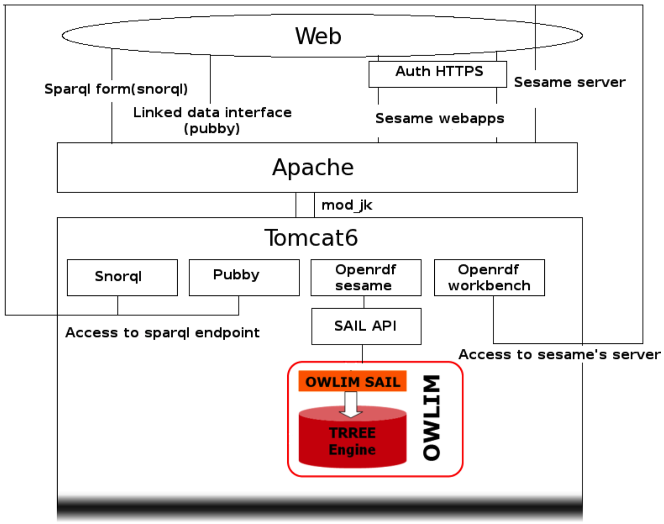

Figure 4: Architecture of the SemUNIT repository

these tools are very useful in some specific cases but they can only transform standard data formats, they does not allow to use a custom ontology (for most of them) and their support and maintenance are often not insured. An handmade XSLT transformation allows a better control of the transformation and does not impose heavy constraints because the SemUNIT ontology is small enough to allow an easy maintenance of the XSL file (when the ontology is modified).

The complete transformation is performed using a python script which executes different procedures before and after the launching of the XSLT transformation. These procedures are the ones which can not be done with XPath functions in XSLT transformation or which need to be done only after all metadata files have been transformed to RDF. The most important of them are:

- transformation of strings belonging to vocabularies into RDF identifiers defined in the SKOS taxonomy described previously (before XSLT);
- transformation of VCard text to RDF FOAF triples (before XSLT);
- management of lom:relation between learning resources. Indeed, relations can either link UNT documents to other UNT documents or to any Web documents (URI HTTP) and URI of UNT documents are created during the transformation. Consequently, this procedure should be done after the whole XSLT transformation.

4. SOME SERVICES

We have designed some services, described below, consuming the triples and taking advantages of the semantic annotations, they are grouped into a Web application[10].

The browsing of metadata in this Web application is insured by Pubby which is integrated via a simple frame in its interface.

In addition to the services described below, we have proposed distinct visualizations: pie charts for values used in metadata elements using controlled vocabularies, timelines...

[10]http://semunt.supelec.fr

4.1 Search Engine

4.1.1 Engine purpose

The engine's purpose is to propose a search service taking advantages of Semantic Web annotations to retrieve learning resources more efficiently. Indeed, most of the users does not know the keywords required to find the documents they need, and even when they know them, useful documents can be lost in hundreds of other documents returned as results.

In addition, a well-designed interface could also be efficient to help users finding information easily. Consequently, a Faceted Search, one of the exploratory Search techniques [11], have been implemented in our engine. In faceted search, facets are conceptual categories allowing to organize data from a large database through a view of clear conceptual groups. There are two kinds of facets : flat and hierarchical [11]. In our search engine, facets have been created to filter results obtained from the classical keyword-based approach and then perform more accurate search. Currently, users can filter results (with facets) on metadata elements taking their values in vocabularies (like structure, format or learning type). For now, only the classification based on the Dewey (mandatory in the UNT context) allows a hierarchical classification.

4.1.2 Detailed description

This search engine works as follows (see figure 5 for an example of a query):

- the user fills in a small form to precise : keywords corresponding to documents they want to find (separated by blank spaces); metadata fields where these keywords are searched: title, description, subject, etc; (optional) person linked to documents they want to find; (optional) role of this person.
- the engine creates a SPARQL query to find these documents, ranks search results and return information about them to the user;
- the user is invited to filter results by selecting elements in different metadata elements presented with faceted search interface;
- the engine returns filtered results and allows the user to change filters and keywords to meet his needs.

The ranking of results is performed using the following algorithm:

- a document gets points each time a keyword appears in its metadata elements;
- the number of points earned depends on the metadata element containing the keywords: 10 for title, 4 for description, 2 for keywords. These numbers have been determined empirically;
- bonus are given when all the keywords are present in the same element (if several keywords has been entered).

Results are sorted according to points they have earned and presented as a list, giving for each document: its title, a link allowing to browse its metadata, a link allowing to get the document and its description and subjects (if their checkboxes are checked).

Figure 5: Search of string "Automatique"

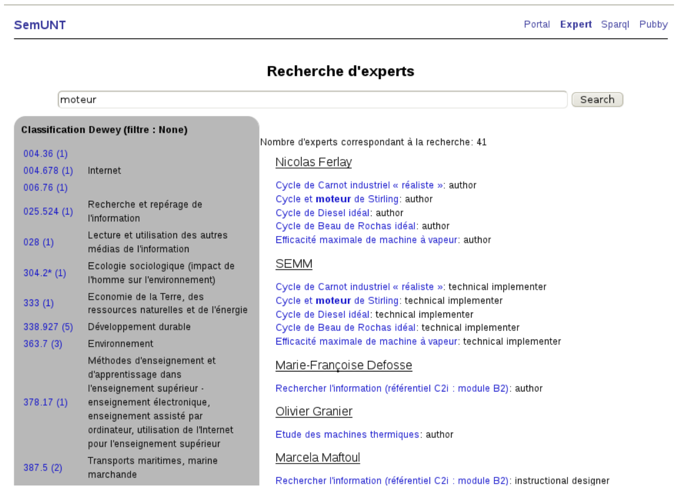

Figure 6: Expert search engine

4.2 Expert search service

This service aims at retrieving people having an expertise in a specific topic (in fact are able to give lessons on a specific topic). The design of this kind of service needs to answer 3 questions :

- which information sources will be used to perform the search;
- which system will be offered to users for choosing the topic(s);
- which metric(s) will be chosen to rank experts found.

Concerning information sources, we used the instances of the class Contribution (which is linked to a contributor and to a kind of contribution). Thus, we can know for each learning resource what kind of contribution a person has brought: author, contributor, publisher, subject-matter expert, etc. The coefficients associated with each kind of contribution have been determined empirically.

About the choice of topic, we have chosen to let users fill in one field form to specify it rather than proposing a list of topics or a list of classification's elements. It offers a more flexible way of specifying topics and allows the reuse of search engine procedures in the ranking system.

5. RELATED WORK

Despite the spread of Linked Data, very few works use the linked data principles in education, the first specific workshop (Linked Learning 2011) was organized in 2011.

The Open University(OU) linked data architectures proposed [15]: a workflow to RDFize data coming from different repositories; an expert search service; a service of students social network; a service allowing students to search and browse free learning materials. This is developed as part of the LUCERO[11] (Linking University Content for Education and Research Online) project. But, the schema[12] used to describe the courses is based on the courseware ontology[13] and on the AIISO ontology[14], two ontologies designed to describe complete courses and internal organizational structure of academic institutions. This schema doesn't include a LOM based ontology, yet widely used.

The mEducator project uses linked data for the publishing of medical educational resources[2]. This project has defined a metadata scheme based on Healthcare LOM (HLOM). It has been designed to be interoperable with existing schemas with the integration of Dublin Core properties as super properties and the use of FOAF properties to describe people. SKOS has been chosen for controlled vocabularies[5]. But the schema used by the mEducator project is a subset of the LOM schema, so some metadata elements like the ones used for the classification are missing. Furthermore, their ontology is only based on RDF(S), so it doesn't take advantage of OWL constructs (like constraints) or OWL tools (like reasoners).

In addition, different services depending on linked data have been proposed : see for instance [12] which proposes an expert search based on linked data metrics or [7] which develops the idea of integrating learning resources in a linked data social network as social object.

6. CONCLUSION AND FUTURE WORKS

This project has shown how the use of Semantic Web technologies coupled with linked data practices could greatly improve e-learning services by giving a formal semantics to metadata elements.

Firstly, we have created the SemUNT ontology which has given a semantic representation to all SupLOMFR metadata elements. This ontology has been designed to be compliant with the meaning of metadata elements and allow an efficient management of vocabularies thanks to the use of SKOS. Among other possibilities, this implementation could ease the management of application profiles vocabularies and the integration of vocabularies in a linked data context. In addition, the ontology has been designed to insured its interoperability with existing ones. Thus, Dublin Core properties have been integrated and FOAF has been chosen to describe persons and organizations.

Then, we have developed tools enabling to automatically transform XML metadata to RDF triples. These tools are based on handmade XSLT transformations and have pro-

[11]http://lucero-project.info/
[12]http://data.open.ac.uk/datasets/
[13]http://courseware.rkbexplorer.com/ontologies/courseware
[14]http://vocab.org/aiiso/schema

duced about 1 million RDF triples from 6456 metadata documents coming from several UNTs: Unit, Unisciel, Uel, Uoh, Uved.

Finally, a SPARQL endpoint[15] has ben proposed and two services[16] have been designed to consume these triples: a search engine using semantic annotations to improve its search efficiency and offer faceted search possibilities and an expert search service allowing to find people skilled in a specific area. These services have shown how a semantic binding of metadata can be used efficiently to improve classical services.

The most original part of this project is the creation of an ontology for the management of educational resources, taking into account the ontology semantics of LOM elements but having a structure (classes and properties) that is not based on that of LOM and reflects the meaning of metadata elements. this ontology is intended to be reused by teams engaged in similar work.

These results and tools will be widely spread in the UNTs community. They have already been presented to representative of different UNTs which have shown their interest. Several scenarios have been imagined as well as creation of multiples services consuming these data and taking advantage of the semantic binding. This project which for now is only a feasibility model, aims to become an institutional project involving all French UNT.

Of course many improvements can be made to our services: search engine and expert search. For example services for documents or expert. Among them: changing the flat facets by hierarchical facets when possible, improving the ranking of the search results, using reasoners reasoners to obtain new information on our data, linking our data to other repositories like DBLP, Semantic Chrunch-Base, Twarql, RDFohloh1, etc.

7. ACKNOWLEDGMENTS

This project has been funded with support from the UNT UNIT.

We want to thank particularly Marie-Hélène Comte and Julia Soyez, librarians in the SemUNIT project.

8. REFERENCES

[1] L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf documents. In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04*, pages 113.3–, Washington, DC, USA, 2005. IEEE Computer Society.

[2] M. Evangelia, D. Taibi, D. Giordano, S. Dietze, H. Yu, P. Bamidis, C. Bratsas, and L. Woodham. Connecting medical educational resources to the linked data cloud: the meducator rdf schema, store and api. In S. Dietze, M. d'Aquin, D. Gasevic, and M.-A. Sicilia, editors, *Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age*, 2011.

[3] A. M. García, S. S. Alonso, and M. A. Sicilia. Una ontología en OWL para la representación semántica de objetos de aprendizaje. In *V Simposio Pluridisciplinar sobre Diseno y Evaluación de Contenidos Educativos Reutilizables (SPDECE)*, 2008.

[4] M. C. Ghebghoub Ouafia, Abel Marie-Hélène and L. Adeline. *Ontology Theory, Management and Design: Advanced Tools and Models*, chapter Building and Use of a LOM Ontology, pages 162–177. 2010.

[5] D. Giordano, I. Kavasidis, C. Spampinato, and P. Bamidis. Developing controlled vocabularies for educational resources sharing: a case study. In S. Dietze, M. d'Aquin, D. Gasevic, and M.-A. Sicilia, editors, *Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age*, 2011.

[6] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology.* Morgan & Claypool, 2011.

[7] E. Kaldoudi, N. Dovrolis, D. Giordano, and S. Dietze. Educational resources as social objects in semantic social networks. In S. Dietze, M. d'Aquin, D. Gasevic, and M.-A. Sicilia, editors, *Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age*, 2011.

[8] D. A. Koutsomitropoulos, A. D. Alexopoulos, G. D. Solomou, and T. S. Papatheodorou. The use of metadata for educational resources in digital repositories: Practices and perspectives. *D-Lib Magazine*, 16(1/2)), 2010.

[9] P. Mohan and C. Brooks. Learning objects on the semantic web. In *IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 195–199, 2003.

[10] M. Nilsson, M. Palmer, J. Brase, et al. The LOM RDF binding-principles and implementation. In *Proceedings of the Third Annual ARIADNE conference*, 2003.

[11] P. Petratos. Informing through user-centered exploratory search and human-computer interaction strategies. *Issues in Informing Science & Information Technology*, 5:705–727, 2008.

[12] M. Stankovic, J. Jovanovic, and P. Laublet. Linked data metrics for flexible expert search on the open web. *The Semantic Web: Research and Applications*, pages 108–123, 2011.

[13] H. Van de Sompel, M. Nelson, C. Lagoze, and S. Warner. Resource harvesting within the oai-pmh framework. *D-lib magazine*, 10(12):1082–9873, 2004.

[14] J. Waitelonis and H. Sack. Towards exploratory video search using linked data. *Multimedia Tools and Applications*, Jan. 2011.

[15] F. Zablith, M. Fernandez, and M. Rowe. The ou linked open data: production and consumption. In S. Dietze, M. d'Aquin, D. Gasevic, and M.-A. Sicilia, editors, *Proceedings of Linked Learning 2011: the 1st International Workshop on eLearning Approaches for the Linked Data Age*, 2011.

[15]http://semunt.supelec.fr/portal/sparql/
[16]http://semunt.supelec.fr/portal/