COMETE – An Educational Search Engine on the Web of Linked Data

Gilbert Paquette, Alexis Miara and Frédéric Bergeron
LICEF Research Center, Télé-université, Montreal, Canada
gilbert.paquette@licef.ca; alexis.miara@licef.ca; frederic.bergeron@licef.ca

***Abstract*— We present here a new Open Educational Resource (OER) repository management tool called COMETE. The evolution of research and practice in the field of OER repositories is moving from OER metadata stored in relational databases to RDF-based descriptions of resources stored in triple stores. COMETE provide the necessary translation from DC, LOM and other metadata schema to an RDF meta-model that offers more intelligent search capabilities, both for designers who are building online environments such as MOOCs, or for students who should be equipped with friendly tools to select resources, activities and co-learners suited to their needs.**

***Index Terms*—open educational resources (OER); resource description framework (RDF); resource repositories; semantic web; web of data.**

I. Introduction

In the last three years, our research on educational resource management moved from Learning object metadata repositories managed by the PALOMA tool [1], to the use of ontology-based annotations within the TELOS system [2] and, finally to the use of semantic technologies for the Web of data [3,4]. The result in a mature tool, COMETE, that is being used in the colleges of Quebec for educational resource referencing and search. The present paper describes the principles and the architecture of the COMETE system and its relation to the ISO-MLR standard [5].

The paper is organized into four sections. In section II, we introduce the notion of an educational resource repository and of a resource manager. Section III presents a recent evolution of e-learning standards, norms and application profiles resulting in the publication of the ISO standard on Metadata for Learning Resources (ISO-MLR) based on the RDF and the Web of data. In the fourth section we provide the principles and the architecture of COMETE, our RDF-based OER manager and we illustrate various kinds of search methods available in COMÈTE.

II. Open Educational Resources Repositories

The term “Open educational resources” was first coined at UNESCO’s 2002 Forum on Open Courseware and defined as “teaching, learning and research materials in any medium, digital or otherwise, that reside in the public domain or have been released under an open license that permits no-cost access, use, adaptation and redistribution by others with no or limited restrictions” Ten year later, UNESCO held in Paris an international OER congress on 20-22 June 2012 where the Paris OER Declaration was issued, recommending that States, within their capacities and authority “foster awareness and use of OER”, “encourage research on OER”, “promote the understanding and use of open licensing frameworks” and “facilitate finding, retrieving and sharing of OER.”

A. First interoperability norms: DC and IEEE-LOM

The idea that educational contents could be seen as "objects" to be reused in multiple contexts dates back to the late 60's but started to become a reality only by the middle of the 90s with the generalization of the Internet [6]. The aim was to insure the reuse of educational objects jeopardized by the diversity of referencing metadata schema around the world.

In 1995, the Dublin Core (DC) metadata initiative proposed a first set of standardized metadata, expressed in XML. Since then, the Dublin Core metadata schema has that was to become one of the most used vocabularies on the Web of Data. In 1996, the IEEE created the Learning Technology Standards Committee to integrate previous work on the concept of Learning Object Metadata (LOM). From then on, major resource repository initiatives bloomed rapidly: ARIADNE in Europe, MERLOT in the USA, EdNA in Australia. These and many other organizations, including ours, joined the GLOBE consortium that operates actually a large repository of nearly one million resources.

B. Potential and Limits of DC/LOM Resource Repositories

Motivations for OER repositories are the growing educational demands in all countries, the limited capacity of face to face education to fulfill the demand in a timely manner, the important effort and cost involved to build online multimedia learning materials and the new possibilities offered by the Internet.

While it is a fact that millions of documents can be found on the Internet using search engines like Google, there is no guarantee that a query will lead to trustable material on which high quality education can be built. On the contrary, OER repositories are maintained by educational institutions and professors providing a certain level of trust for the quality of referenced resources and giving precious information to users, that helps make more focused queries. These query prevent

blind search based on vague keywords that leads to thousands of references that one needs to read to understand what kind of content they provide. Finally, the vast majority of these learning objects are in the public domain to be reused free of charge. These resources can be adapted or aggregated, and referenced back in a repository to extend the availability of good learning material.

After a decade of research and practice in this field a number of limitations to a larger use of OER repositories still exist. The most important ones are the multiplicity of norms and Applications profile, as well as the imprecision of most proprietary metadata schema.

III. ISO-MLR AND THE WEB OF DATA

Although the Dublin Core and the IEEE-LOM are widely used to describe learning resources, interoperability among metadata sets from multiple repositories is still challenging.

For example, instead of using ISO 8601, a DC Date element can be written in plain language making it impossible its processing by queries. Ambiguous definitions pose another challenge. For example a Date element can represent a resource creation time, a time of update or a time of publication. As mentioned above, LOM records can be based on a wide variety of Application profiles each defined in their own way by various organizations.

A. *ISO-MLR: an OER referencing standard based on RDF*

The ISO/IEC 19788 standard [5], in short ISO-MLR, is intended to provide optimal compatibility with both DC and the LOM. It insures the coherence and the non-duplication of concepts by proposing an RDF-based data model. It prevents the proliferation of non-interoperable application profiles. It supports the extension of description vocabulairies in precise ways while preserving interoperability as well as multilingual and cultural adaptability requirements from a global perspective, while integrating ressource referencing and search with other data sets in the Web of linked data.

The graph in Fig.1 shows part of the ISO-MLR RDF model. The ovals represent classes of resources, the rectangles are value types, and properties are written over the links. This graph summarizes the RDF triples in section 5 of the standard. Here are some of the triples present on Fig.1:

(Learning resource, has learning activity, Learning activity)
(Learning activity, learning method, *method value*)
(Learning resource, has contribution, Contribution)
(Contribution, has contributor, Person)
(Annotation, annotation date, *date value*) ….

Links like the second and last one represent object properties linking various kind of resources, while the others are data properties specifying values of properties.

B. *A standard for the Web of data*

ISO-MLR, uses technologies like RDF and RDF schema, integrates well to a Web of linked data. The origin of the Web of data, also termed “Semantic Web”, dates back to 2001 when the actual director and founder of the Web, Tim Berners-Lee and his colleagues proposed to extend the Web to use URI’s to represent not only pages of information but also people, real-world objects and also abstract concept and properties. These entities and the values of their properties can be linked together by declaring RDF triples.

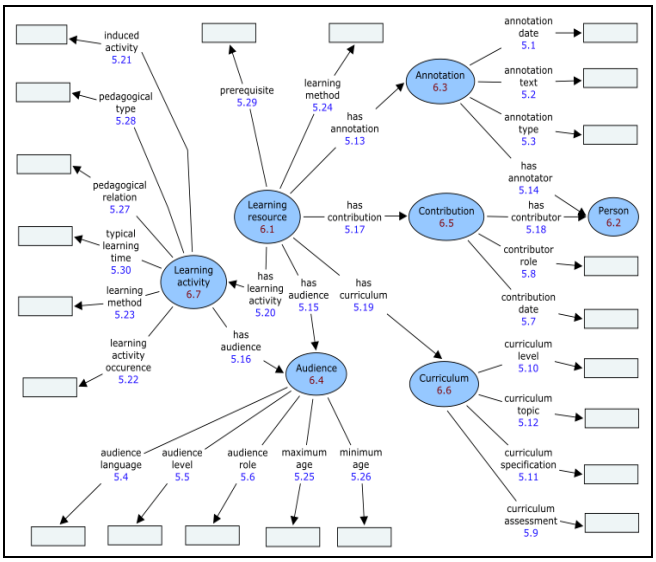

Fig. 1. Part of the ISO-MLR RDF model

It then becomes possible to describe the semantic of Web pages beyond the syntax of natural languages and their inherent ambiguity. A Web of linked data enables computer agents to follow the links and perform more intelligent operations using the knowledge behind the words.

For example, the SPARQL Query Language [7] enables queries within the huge graph of RDF triples that constitutes the Web of linked data. At the end of 2014, this graph grouped 74 billions RDF triples from 353 datasets. And it is still growing. Within this graph, the DBpedia dataset contains a large part of the information in Wikipedia, while the FOAF dataset provides information about persons having a URI on the Web. Terms in a vocabulary (concepts and properties) are linked with terms in another vocabulary. For example, “persons” in DBpedia is related to “persons” in FOAF and their geographical localization can be found in another vocabulary or dataset such as GEONAMES.

In the same way, terms in ISO-MLR are linked to terms of other vocabularies on the Web of data. For example, iso-mlr5:Person in the graph of Fig.1 has the same meaning as foaf:person or dcterms:person. This means that a computer agent that would search for an iso-mlr:learning_resource can also ask for its iso-mlr:Contributors, find these persons and retrieve their Wikipedia pages from DBpedia, their email from FOAF and their localization from GEOBASE.

IV. COMETE, A RDF-BASED RESOURCE MANAGER

COMETE is a learning resource repository manager based on the RDF approach. It allows locating, aggregating and retrieving educational resources that constitute the heritage of

an organization. Basically, it is a database containing metadata about learning resources on which users can perform queries to find and discover educational material that they can reuse for their various needs.

Fig. 2 describes the technical architecture of a COMETE implantation instance. It's a 3-tiers client-server architecture developed in Java. Various web applications powered by an Apache Tomcat server provide specialized REST services that allow different types of clients to exploit the open data contained in the repository. Most of the clients use their favourite web browsers to access the system through a user-friendly web interface. A SPARQL endpoint is also available for advanced users who want to directly access the raw RDF triples to build various applications or Web services.

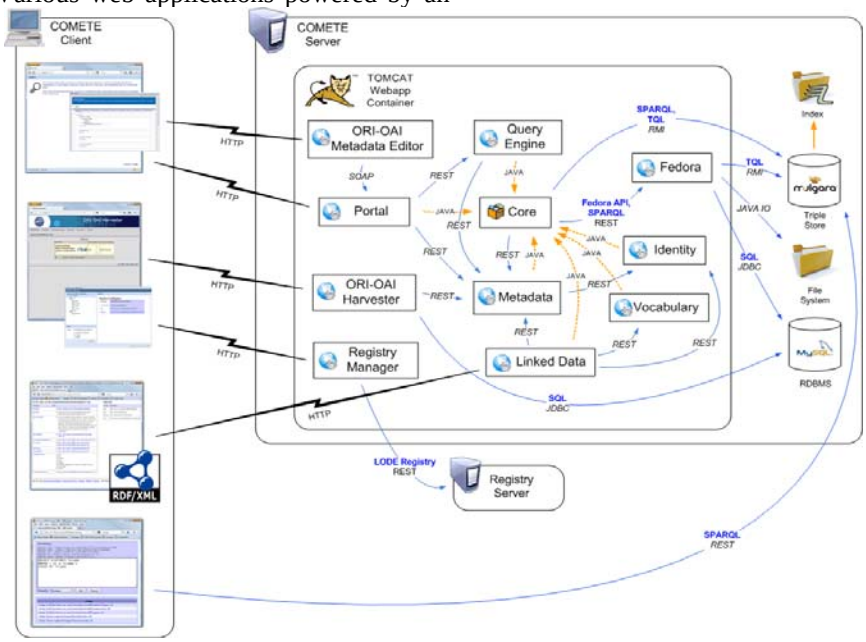

Fig. 2. COMETE architecture schema

A. *Integrating new resources in the triple store*

The integration of resources inside a COMETE repository is done by imports of their metadata records. The metadata records can be imported manually by uploading an archive file containing a collection of metadata records. Most of the time, however, either an OAI-PMH Harvester or a HTML Spider will harvest the metadata records automatically. In such a case, a Harvest Definition will declare the technical information required to access the repository to be harvested. It is also possible to program harvest schedules so that the process is executed periodically to make sure that new or updated metadata records are always imported to the system.

These records are ingested by the system and a XSL transformation extract data for generating all pertinent triples. COMETE enables data mining across multiple metadata schemas like Dublin Core, IEEE LOM and other application profiles. The result of this process is a homogeneous graph of data in accordance with COMETE's internal meta-model (partly shown on Fig. 3).

All the triples that are generated are stored into Mulgara [8], an open source RDF triple store system, where data is organized around various RDF graphs. The default graph contains all the triples about learning resources whereas some other specialized graphs manage SKOS thesaurus and other different views of the system.

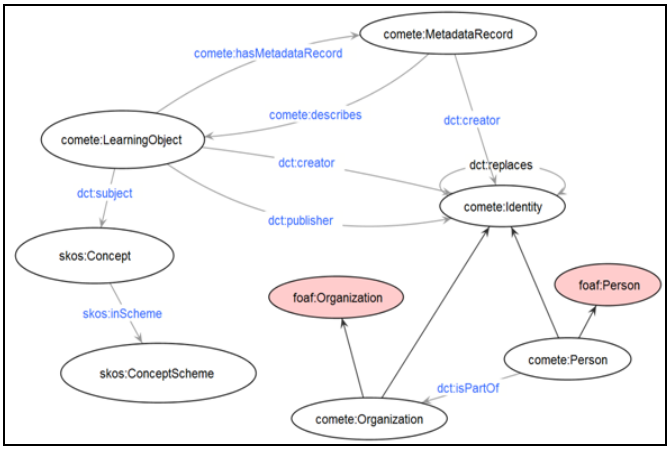

Fig. 3. COMETE metamodel (main classes)

As a semantic network, the RDF graph represents the model entities as nodes. Mains nodes are learning resources (Learning Object), persons and organizations (Identity) and element of vocabulary (SKOS Concept).

The Identity module showed on Fig. 2 implements the management of metadata about persons or organizations. This includes importation of identities, identity resolution of data that represents the same person or organization, making sure it stays unique, and completing it as new details are known. Furthermore, manual merge of identities is also provided within a set of administrative tools for a better control of data integrity.

The Vocabulary module (on Fig. 2) implements the management of vocabularies and thesauri imported from VDEX or SKOS formats, unambiguously identifying the vocabulary that a term is from, converting from one format to another, replacing a vocabulary when updates are available, publishing vocabularies automatically and providing user interface elements reusable by other modules, such as efficient vocabulary term choosers for queries to the repository.

This module manages also SKOS concept alignment between different ontologies (or vocabularies) that is taken into account by the query engine. For example the mapping between different school-level taxonomies in different countries enables one to search resources for Junior High School in United States, and the results may contain pertinent Secondary School I-III tutorials produced in Québec.

B. *Querying the triple store*

All of the previously presented modules provide rich graphs of data that allow processing more "intelligent" searches in the repository than before.

COMETE provides four search modes. The simple mode only needs a field of keywords, as in a Google search. The CERES implementation of COMETE (Fig. 4), now in use in Colleges of Québec grouping over 40 000 resources, will serve to illustrate the other modes.

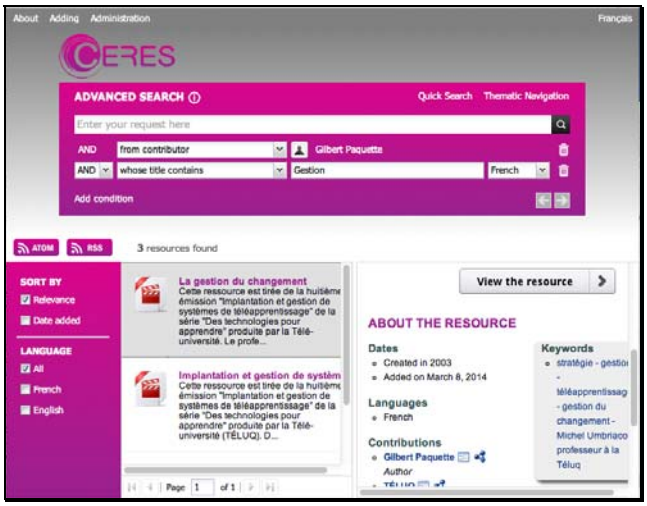

Fig. 4. COMETE Advanced Search Interface

Queries like the one of Fig.4 are translated in SPARQL language by the Query Engine module (shown on Fig.2) and then run on the triple store to extract the information. Suppose now we seek the resources authored by a contributor dealing with the subject of change management ("gestion du changement" in French). Fig 4 illustrates how to fill such a query. The user could add other properties to refine the search. The right part of the figure shows the properties of a selected resource. Links to the author's and his organization's information, and also taxonomies of concepts are clickable to navigate on the global RDF graph.

A third way to search resources inside a repository is to use the Thematic Navigation mode that makes use of the Linked data module (also shown on Fig.2). It lets user discover resources directly from the categories in the available vocabularies. Results are returned and displayed to the user interface using the alignment of vocabularies integrated in the system; for example, queries may be extended with "include equivalent categories from the Library of Congress".

Finally, a fourth search mode in COMETE user interface is the Collection mode. It offers to users a list of preset complex queries to avoid having to enter them by hand. For example, one could ask the system: "give me all the resources on Algebra and non-Euclidean geometry produced last month from authors at the Université de Montréal".

C. *Linking with the Web of data*

The link with the web of data, as a global data space, is achieved by respecting the basic principles of Linked Data: representing all entities by HTTP URIs, dereferencing URIs over the HTTP protocol into a description of the identified object or concept served in different versions: HTML page for web browser clients, RDF/XML for software agent. A COMETE vocabulary details the class and property definitions in the meta-model, reusing existing vocabularies such as Dublin Core, FOAF and SKOS. The publishing of data via a SPARQL endpoint allows interaction with COMETE data by external systems.

V. Conclusion

We have presented a solution to one of the main problems in Open Educational Resources repositories, which is the multiplicity of norms, standards and application profiles that preclude efficient search for resources within multiple repositories. We have built a Linked data OER repository manager, COMETE, relying on semantic web techniques, largely compatible with the new ISO-MLR standard.

References

[1] Paquette G, Lundgren-Cayrol K, Miara A, Guérette L (2004) The Explor@2 Learning Object Manager, in R. McGreal (ed), Online education using learning objects. pp 254-268. London: Routledge/Falmer.

[2] Paquette, G. (2010). An ontology-driven System for e-learning and knowledge Management. In Paquette, G., Visual Knowledge Modeling for Semantic Web Technologies: Models and Ontologies. Hershey, PA: IGI Global, pp 302-324

[3] Allemang D. and Hendler J. (2011) Semantic Web for the Working Ontologist – Effective Modeling in RDFS and OWL. 2nd Edition. Morgan-Kaufmann/Elsevier, Amsterdam.

[4] Heath, T. et Bizer, C. (2011). Linked data: Evolving the web into a global data space. In Synthesis Lectures on the Semantic Web: Theory and Technology, 1(1), 1-136.

[5] ISO-MLR (2013) ISO-IED 19788 Information technoogy – Learning, education and training – Metatda for learning resources multipart standard. http://en.wikipedia.org/wiki/ISO/IEC_19788.

[6] Duval, E. and Robson, R. (2001) Duval, E. and Robson. R. Guest Editorial on Metadata. Interactive Learning Environments, Special issue: Metadata, Volume 9-3, December 2001, pp. 201-206

[7] SPARQL – SPARQL 1.1 Query Language, W3C Recommendation, 21 March 2013. http://www.w3.org/TR/sparql11-query/

[8] Mulgara RDF Triple Store System. http://en.wikipedia.org/wiki/Mulgara_(software)