LARGE-SCALE STORAGE AND RETRIEVAL OF EDUCATIONAL METADATA USING AN RDF STORE

Xavier Ochoa
Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
xavier@cti.espol.edu.ec

Andre Ortega
Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
andre.ortega@cti.espol.edu.ec

Gladys Carrillo
Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
gladys.carrillo@cti.espol.edu.ec

Carlos Villavicencio
Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador
carlos.villavicencio@cti.espol.edu.ec

Abstract: There is a long tradition of the use of metadata stores in the educational setting. These stores usually present a mismatch problem between the structure of the metadata, usually some form of XML, and the structure of the repository itself, for example relational databases, or document-oriented stores. These mismatches lead to reduced functionality of the metadata store. This work presents a different alternative, using RDF as the internal representation of educational repositories. The architecture and actual implementation of such a system is also discussed. The use of a semantic representation of the educational metadata opens the door for novel functionalities that could provide a more intelligent repository to the final user.

Keywords: Learning Object Repositories, RDF, Semantic Web

1. Introduction

The concept of Learning Object has evolved from the need to reuse digital learning materials. Learning Object Technologies offer economic as well as pedagogical advantages. The learning materials are created just once, but used several times in different contexts, compensating the high cost of production. Also, high quality, thoughtfully designed, multimedia materials could be easily accessed by any instructor or learner.

Learning Objects can be shared in several ways. They can be just published on the web, made available in online forums or even pass personally from user to user. This work however, concentrates in the most formal way of learning object sharing: Learning Object Repositories. To share an object in this way, the object is indexed in what is called a Learning Object Repository (LOR). In their most common form, LORs usually store the learning object itself and the metadata instances associated with it. These LORs provide some sort of indexation facility, where users can add new learning objects together with their metadata. Also, some sort of search or browsing facility is provided to provide access to the content of the repository.

Any data that can be used to describe a learning object can be considered as learning object metadata. According to the IEEE Learning Technologies Standard Committee, the purpose of the metadata is to facilitate the "search, evaluation, acquisition, and use of learning objects". Therefore, a general definition of learning object metadata is any piece of information that can be used to search, evaluate, acquire and use learning objects. For example, the title of a learning object would help to find a relevant learning object. A review created by a user would help to evaluate the relevance of the object for another user. The link pointing to actual resource, as well as the information about the copyrights of the object would help to acquire the object properly. Finally, the technical information about the object, such as file type or size, would help the user to select the right tools to use the object. The most commonly used metadata standard for learning objects is LOM (Learning Object Metadata) (IEEE, 2002). This standard was sanctioned by the IEEE Learning Technologies Standard Committee. LOM proposes around 50 different metadata fields grouped into nine categories.

All these definitions and standards have been used to create computational systems that could help to share and reuse educational materials. For example: the ARIADNE Knowledge Pool System (Duval et al., 2001), Connexions (Baraniuk, 2007), MERLOT (Malloy and Hanley, 2001) and INTUTE, among others. All these systems have been operational for many years and had fulfilled their technical purpose. However, their limited size (Ochoa and Duval, 2008) has prevented them to be really useful for their end-user: teachers and learners. To solve the size issue, the biggest LORs have formed the Global Learning Object Brokered Exchange (GLOBE). The main objective of GLOBE is to enable the sharing of learning objects between repository networks worldwide. In order to reach that goal, each member adheres to a set of technical standards that facilitate the interoperability between repositories. The first step to achieve interoperability is to be able to communicate and to obtain the metadata from partner repositories. Currently the Simple Query Interface (SQI) is the selected standard to obtain metadata through federated queries (Ternier, 2008). Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH), on the other hand, is the standard used to harvest the metadata from other repositories (Van de Sompel et al., 2004). With these two standards, GLOBE members are able to obtain the metadata describing learning objects stored in partner repositories.

The creation of GLOBE, however, has revealed several problems that were not apparent in the isolated LOR setting. This work presents such problems in Section 2, proposes a solution based on a Distributed RDF Store in Section 3 and sketches a design of such solution used to implement the main Educational Metadata Repository inside a real project. This work finishes with some conclusions of the lessons learnt during the implementation of this RDF Repository for educational metadata.

2. Problems with current LORs

Due to their initial design, when current LORs are used to support large-scale learning solutions they present serious functional, scalability and interoperability problems. This section discuss these issues and their causes:

There is much more to Learning Objects than LOM

While the main purpose of the creation of LORs was to provide a specific way to store the Learning Object metadata, this constraint heavily limit their usefulness in real-world, complex and large-scale learning solutions. Such solutions often deal with different type of entities related, but not equal, to learning objects. For example, it is useful for a learning solution to store a user profile. This profile is basically metadata about the entity user. This entity is very related with the learning object: the user download a learning object, the user rate a learning object, the user shared a learning object with another user, etc. Other example is Learning Paths, sequences of Learning Objects. Learning Paths could exist or could be created and need to be stored similarly to Learning Objects, specially the relation to the different learning objects that compose them.

The current solution to this problem is to create separated repositories, databases or tables to store the information about all the different entities that take part in the learning solution. Moreover, special databases or tables are created to store their relationships that usually many-to-many: A user can download several Learning Objects, a Learning Object can be downloaded by many users, etc. This solution-specific repository or database makes very difficult to reuse it across learning solutions.

Databases are not suitable for LOM

LOM is a complex hierarchical structure designed to store several values for a single field or element. Naturally LOM maps into hierarchical formats, such as XML. However, most of the LOR implementations are based on Relational Databases. This imposes a costly mapping from the external representation of the information (LOM) and the internal representation of the same information (Relational Tables) (Florescu and Kossmann, 1999). This complex translation often results in low-performing insertion and retrieval operations. While the number of LOMs stays low, this hit in performance could be acceptable, however, in large-scale implementations (1 million or more LOMs), the time to answer to a query could be easily in the order of seconds.

Several solutions has been tried to solve this problem. The first one is to reduce the complexity of LOM limiting its features. For example, a Learning Object could have just one title or no Annotations are stored. This again, could work in an isolated environment where the Metadata is homogeneous, but in a heterogeneous system such as GLOBE, it is impossible to determine what parts of LOM will be filled by the different communities. A second approach to reduce the complexity is to use an XML database to eliminate the need to map from LOM to Relational Tables and vice versa. However, XML databases are considerable slower than Relational Databases (Nicola and John, 2003), defeating the original purpose of their use. Finally, the most successful solution to date is the use of document-based databases, such as Lucene (Hatcher et al., 2004) to create a text index of the text contained in the XML. This approach accelerates considerably the query-time, but eliminate the hierarchical structure of LOM because such document databases posses only a flat representation of the information.

Not everyone speaks LOM

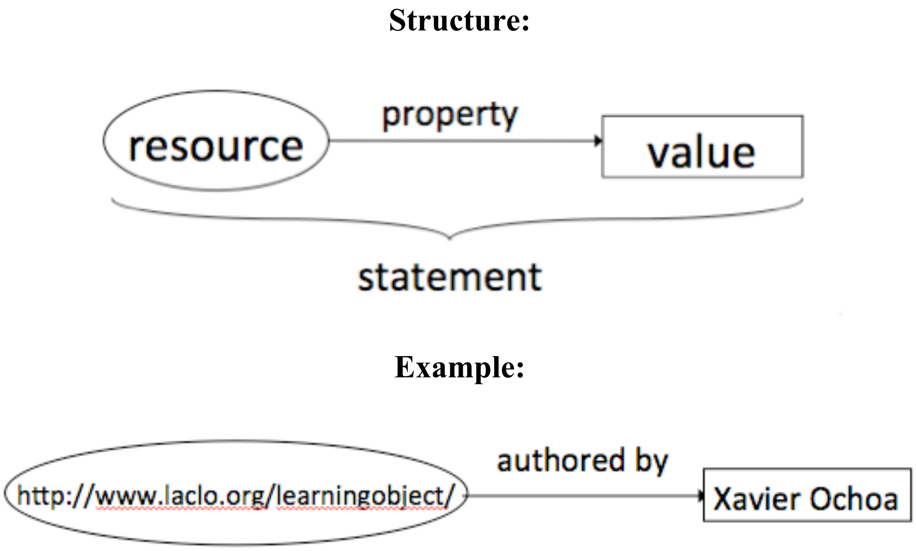
While successful, LOM is not always used to represent learning materials. Competing standards such as Dublin Core (DC) or MPEG-7 are sometimes used to represent materials that could be used for learning. The main solution to integrate this kind of objects, which does not have a LOM description, has been to use metadata translation. This translation, however, is not perfect and information is lost. This lower-denominator approach hinders interoperability among repositories that could be part of a large-scale learning solution.

It can be noted that the origin of this issues is not the original idea of the LOR, but the use of LOM, and more specifically, the XML binding of LOM as the core of the LOR. The next section proposes the use of RDF as the core of a more scalable, functional and interoperable type of LOR.

3. Proposed Solution: RDF Stores

The Resource Description Framework (RDF) is a W3C standard format for describing resources in the WorldWide Web and elsewhere. RDF is part of the Semantic Web vision that aims to improve web information with valuable meaning to be shared and processed by humans and machines. In this context, RDF is the language used in Semantic Web to express metadata statements.

The structure of an RDF statement is called a triple and consists of a *subject,* a *predicate* and an *objec*t, also expressed as *resource, property, value*. For example, the learning object could be the resource, identified by its location (URI). The statement that will be expressed is that it has ben authored by a person, so the property will be "authored by". Finally the value of will be the identification of the author. Figure 2 expresses that: Xavier Ochoa is the author of the resource identified by *http://www.laclo.org/learningobject/*



Structure:

Example:

Figure 1. Resource Description Framework (RDF) Structure and an example

RDF can be used to express any metadata field as a statement. The object been described is the resource, the metadata field been described is the property and the metadata value is converted in the RDF value. The statements could also represent hierarchies and multiple values very easily, making it ideal for LOM. Moreover, the values of RDF statements could be resources themselves, enabling the homogenous description of heterogeneous entities, one of the main problems of the current LORs based on LOM. The collection of statements forms a graph. This graph can use a unique storage to keep all related data to learning environments such as users, courses, lessons, learning objects, etc. in a more efficient way where all the entities could be linked to other internal entities and also to external resources.

Table 1 presents an excerpt of the triple representation of a course metadata. It can be seen how information about lessons is described as Uniform Resource Identifiers (URI), with these links the information about the lessons can access and also could link to different resources in the same way.

Table 1. RDF representation of a Course

Resource	Property	Value
http://igualproject/COURSE/ID#1	http://igualproject/title	Fundamentals of Java
http://igualproject/COURSE/ID#1	http://igualproject/author	xochoa
http://igualproject/COURSE/ID#1	http://igualproject/language	en
http://igualproject/COURSE/ID#1	http://igualproject/lesson	http://igualproject/LESSON/ID#35
http://igualproject/COURSE/ID#1	http://igualproject/lesson	http://igualproject/LESSON/ID#36

Using a RDF store as a basis of Learning Object Repository of large-scale learning solutions solve the problems identified in the previous section:

> *There is much more to Learning Objects than LOM:* An RDF store is designed to natively handle unlimited number of different entities, being them Learning Objects, Learning Paths, Students Profiles, Courses, User Actions, etc. All of those entities will be transformed into Resources and all the relevant information about them will be coded as properties and values. The relation between entities could be easily managed joining two resources through a property.
>
> *Databases are not suitable for LOM:* RDF Stores could handle all the complexities of LOM in a very elegant way. There is no information is lost in the transformation from LOM-XML to LOM-RDF. While full-text search could still be an issue, given than RDF Stores are specialized in recover objects through linking, a Lucene index could be use to support this functionality while the RDF Store could manage all other types of query.
>
> *Not everyone speaks LOM:* RDF Stores are metadata standards agnostic. Any type of metadata standard could be converted to RDF. Moreover, if there are descriptions about a same resource in two different formats, they could be easily combined in the store. Also, metadata instances created to describe different entities could be linked through the RDF graph.

The use of a RDF store seems to be the natural evolution of LORs given the more natural adaptation to their new role as center of large-scale learning solutions.

4. Case Study: Architecture of the IGUAL Project

The Innovation for Equality in Latin American Universities (IGUAL) Project proposes the use of innovative learning technologies to help university students from public schools to bridge the knowledge and skill gap with their private schooled counterparts. It will provide a tool where teachers can create courses, lessons with different learning activities and all these materials will keep in a repository where users can access and use in a suitable way.

One of the main components of the IGUAL Project is the Repository and Recommender Service. This component has been implemented using a distributed RDF Store in its nucleus. The RDF Store selected for this task was 4Store (Harris et al, 2009). The selection was based on its scalability and distributed nature.

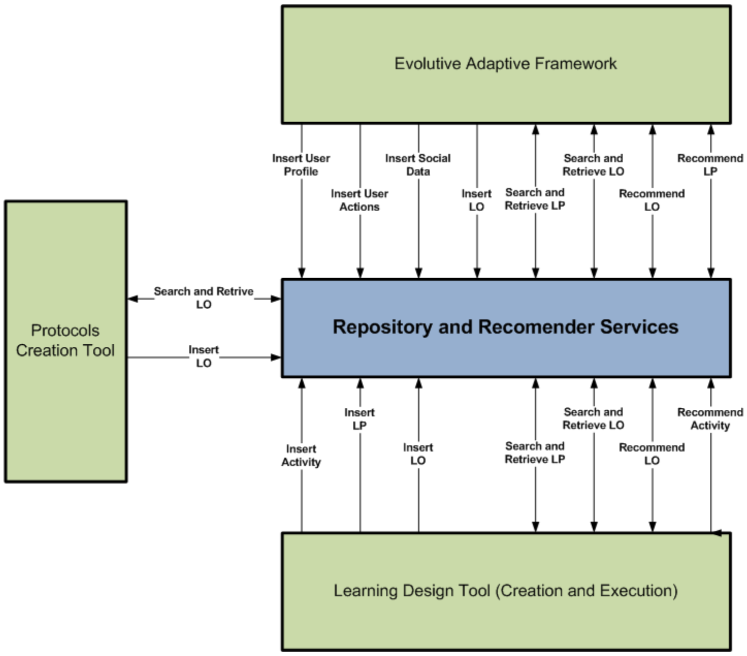


Figure 2. Communication of the Repository and Recommender Service with other components of the IGUAL architecture

To work, the IGUAL Project need to manage information about different entities: Users, Courses, Lessons, Objectives, Learning Activities, User Actions, Learning Paths and Taxonomies mainly. The Repository is responsible for storing this information and retrieving it base on the needs of the rest of the components. Figure 3 shows the simplified RDF graph of this store where only the relationships are highlighted.

In a traditional storage using relational databases, for each entity a schema must be defined with fixed fields. RDF can be used to describe metadata in a simpler and more efficient way. The Metadata Repository of IGUAL Project uses this standard for the storage of entities used in the learning solution.

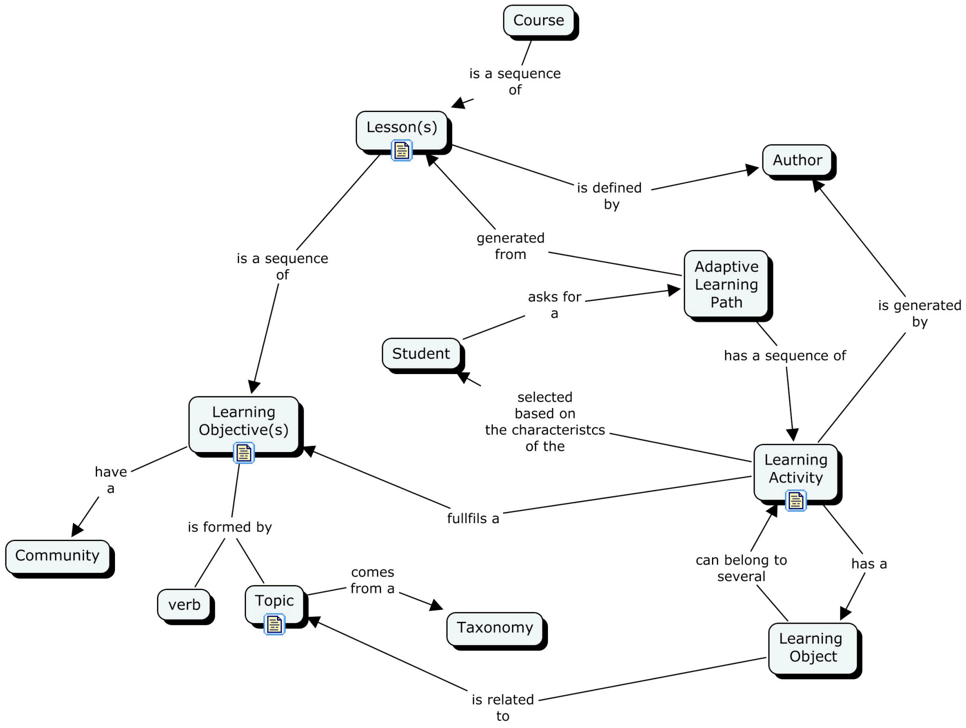


Figure 3. Simplified RDF Graph of the IGUAL Project

When the repository receive a request to save an entity, its metadata is represented as a collection of triples, for example:

Course Metadata in xml format

```
<course>
        <courseID>1003</courseID>
        <title>Some course</title>
        <description>A brief description of the course /description>
        <language>en</language>
        <author>xochoa</author>
</course>
```

Course Metadata in triples

resource	property	value
1003	title	Some course
1003	description	A brief description of the course
1003	language	en
1003	author	xochoa

Then, these triples are stored in the RDF database. The resource and property are translated into URIs:

http://igualproject.org + / + entity name + / + ID# + resource

http://igualproject.org + / + entity name + / + KEY# + property

Course Metadata in RDF Store

```
<http://igualproject.org/COURSE/ID#1003> <http://igualproject.org/COURSE/KEY#title> "Some course"
<http://igualproject.org/COURSE/ID#1003> <http://igualproject.org/COURSE/KEY#description> "A brief description of the
course"
<http://igualproject.org/COURSE/ID#1003> <http://igualproject.org/COURSE/KEY#langugage> "en"
<http://igualproject.org/COURSE/ID#1003> <http://igualproject.org/COURSE/KEY#author> "xochoa"
```

When the repository received a query request, a SPARQL (SPARQL Protocol and RDF Query Language) sentence is executed in the RDF database. For example:

To get course information with id=1003

```
SELECT ?p ?o FROM <http://igualproject.org/COURSE> WHERE
{<http://igualproject.org/COURSE/ID#1003> ?p ?o}
```

The database response is then internally translated in a friendly response like xml format. This response is then transferred to the component that required.

To provide easy access to the store an Application Programming Interface (API) is provided to the rest of services of the IGUAL Project. Part of this API could be seen in Figure 2.

5. Conclusions

From the experience of IGUAL Project, it can be concluded that it is possible to build a fully functional repository for educational metadata using just a RDF store. This RDF store solve most of the problems that traditional LORs has with the storage of LOM and related metadata. This RDF Store is more efficient and simple to use than a traditional storage since it automatically provide linkage between different entities and provide the flexibility to add new entities or change the metadata formats any time. Additionally, the use of the RDF store provides a simple way to connect the repository with external resources, converting the repository in a node in the Open Linked Data network (Bizer et al, 2009).

Acknowledgment

We gratefully acknowledge the collaboration of colleagues in the IGUAL project, as well as the financial support from European Union (DCI-ALA/19.09.01/10/21526/245- 315/ALFAHI (2010)123) of the ALFA III program). Although this article has been produced with the financial assistance of the European Union, the contents of this article are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the position of the European Union.

References

1. Baraniuk, R. G.Iiyoshi, T. & Kumar, M. S. V., ed., (2007), Opening Up Education: The Collective Advancement of Education through Open Technology, Open Content, and Open Knowledge, MIT Press, chapter Challenges and Opportunities for the Open Education Movement: A Connexions Case Study, pp. 116--132.

2. Bizer, C.; Heath, T. & Berners-Lee, T. (2009), 'Linked data-the story so far', International Journal on Semantic Web and Information Systems (IJSWIS) 5(3), 1--22.

3. Duval, E.; Warkentyne, K.; Haenni, F.; Forte, E.; Cardinaels, K.; Verhoeven, B.; Van Durm, R.; Hendrikx, K.; Forte, M.; Ebel, N. & others (2001), 'The Ariadne knowledge pool system', Communications of the ACM 44(5), 72--78.

4. Florescu, D. & Kossmann, D. (1999), 'Storing and Querying XML Data using an RDMBS.', IEEE Data Engineering Bulletin 22(3), 27--34.

5. Harris, S.; Lamb, N. & Shadbolt, N. (2009), 4store: The design and implementation of a clustered rdf store, in '5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009)', pp. 94--109.

6. Hatcher, E.; Gospodnetic, O. & McCandless, M. (2004), Lucene in action, Manning Publications.

7. IEEE (2002), 'IEEE 1484.12.1 Standard: Learning Object Metadata, http://ltsc.ieee.org/wg12/par1484-12-1.html, retrieved 2/04/2007'.

8. Malloy, T.; Jensen, G.; Regan, A. & Reddick, M. (2002), 'Open courseware and shared knowledge in higher education', Behavior Research Methods, Instruments, & Computers 34(2), 200--203.

9. Nicola, M. & John, J. (2003), Xml parsing: a threat to database performance, in 'Proceedings of the twelfth international conference on Information and knowledge management', pp. 175--178.

10. Ochoa, X. & Duval, E. (2008), Quantitative Analysis of Learning Object Repositories, in 'Proceedings of the ED-MEDIA 2008 World Conference on Educational Multimedia, Hypermedia and Telecommunications', AACE, Chesapeake, VA, pp. 6031-6048.

11. Ternier, S. (2008), 'Standards based Interoperability for Searching in and Publishing to Learning Object Repositories', PhD thesis, Katholieke Universiteit Leuven.

12. Van de Sompel, H.; Nelson, M.; Lagoze, C. & Warner, S. (2004), 'Resource Harvesting within the OAI-PMH Framework', D-Lib Magazine 10(12), 1082--9873.