Towards the use of Semantic Learning Object Repositories: Evaluating Queries Performance in two Different RDF Implementations

Henrique L. dos Santos, Gladys Carrillo, Cristian Cechinel and Xavier Ochoa

***Abstract*— Learning Object Repositories (LOR) are an essential component of the e-Learning ecosystem and have being normally serving the purpose of cataloging, storing, retrieving and delivering Learning Objects to be used inside e-Learning applications. Next generation of LORs needs to overcome some major shortcomings current LORs present and involve other entities that are part of the e-learning process (teachers, students, lessons, courses, activities, learning paths, etc.) in a way that they are all fully integrated and linked-up. Semantic web technologies such as RDF are the natural choice to implement these requirements into Semantic Learning Repositories. One important factor limiting the implementation of this kind of systems is the uncertainty about their performance. The present paper describes an initial study that compares the performance of two distinct RDF native database implementations (4store and Jena Apache) in the specific context of a Semantic Learning Repository. The performance tests were run to evaluate two different aspects of the databases implementation: the time to upload the RDF data to the databases, and the response time for running the queries. The results showed that 4store performed better than Apache Jena for all the scenarios we evaluated.**

***Index Terms*— performance analysis, RDF database, Semantic learning object repositories**

I. INTRODUCTION

LEARNING Object Repositories (LOR) have been the backbone for the construction of e-learning systems that provide access to a large amount of learning resources. Traditionally, these LORs have been implemented as document repositories, that is, they are centered only on one entity, in this case, the Learning Object. The information stored in a traditional LOR is the learning resource file and the metadata, in a predefined format, describing that resource. In the case of the learning resource file, some LORs store only a reference to where the file is stored and these LORs are called "Referatories". The traditional design of LOR while useful for the direct retrieval of Learning Objects, present several shortcomings when used in real-life e-learning systems. First, e-learning systems should manage much more diverse entities than just the learning objects. The learner, the teacher, the lesson (sequence) should also need to be taken into account. E-learning systems usually solve this shortcoming having several repositories for different type of entities: one for the learning objects, other of the user profile and another for the lesson structure. While this let the e-learning system to store all the needed information, it adds complexity to the system and makes very difficult to maintain the very necessary relationships (links) between entities [1]. A second major shortcoming of traditional LORs is their reliance on a single metadata format to describe the learning resources. In the best-case scenario this format will be a standard such as Learning Object Metadata (LOM) or Dublin Core (DC), otherwise, it will be an ad-hoc structure. Due to this reliance on a single metadata format, a whole area of research on Learning Object Interoperability has been developed in order to be able to interchange information between several repositories [2]. These interoperability issues, again, add complexity to the design of e-learning system, especially if it is desired that their data remain open for others to be used.

Finally, being based on predefined formats for their metadata, traditional LORs are designed to operate with a static structure. If new elements or entities are added to the e-learning system, the LOR will be unable to accommodate them and a new repository, or a major re-design, will be needed to store their information. Rapid changing and adaptable e-learning systems could only communicate with LORs as a source of information, but not as an integral part of the architecture of the system [3]. All these shortcomings demand a drastic redesign of the concept of Learning Object Repository to be the main persistence component of modern e-learning systems.

The concept of Semantic Learning Repositories solves the

This work was supported by CYTED (Ibero-American Programme for Science, Technology and Development) as part of project “RIURE - Ibero-American Network for the Usability of Learning Repositories”, code 513RT0471 (www.riure.net) and by FAPERGS (Research Support Foundation of Rio Grande do Sul – Brazil) through Edital 02/2014 - PROGRAMA PESQUISADOR GAÚCHO – PqG.

H. L. dos Santos is with the Computer Engineering Course, Federal University of Pelotas, Brazil (e-mail: henriquelds94@gmail.com)

C. Cechinel is with the Faculty of Education, Federal University of Pelotas, CEP 96010-020 Brazil (phone: +55-53-39211431; e-mail: contato@cristiancechinel.pro.br).

G. Carrillo is with Faculty of Electrical and Computer Engineering at the Escuela Superior Politécnica del Litoral (ESPOL), Ecuador (e-mail: gladys.carrillo@cti.espol.edu.ec)

X. Ochoa is with Faculty of Electrical and Computer Engineering at the Escuela Superior Politécnica del Litoral (ESPOL), Ecuador (e-mail: xavier@cti.espol.edu.ec)

LOR shortcomings mentioned here. First, if all needed information can be stored in a single repository, there is no need for the e-learning system to include or connect with other types of repositories. Different types of e-learning systems could include different description for the entities and even different entities depending on the learning process they are supporting. Second, the use of Semantic Technologies leads to a format-free repository. Any metadata standard could be used to describe the existing entities. Mapping between metadata standards or ad-hoc structures is greatly facilitated by the use of RDF triplets to store information. The interoperability issues are also reduced if the data is published as Open Linked Data [4]. In this way, it can be easily consumed by any other Semantic Learning Repository or e-learning application. Finally, changes in the metadata formats and/or stored entities can be easily incorporated into the Semantic Learning Repository without need to change its functionality.

Given the internal structure of the data and the flexible nature of RDF implementation, the natural choice for the implementation of a Semantic Learning Repository is to use an RDF store. However, the perceived difference in performance between traditional Relational Database Management Systems (RDBMS) and RDF stores has been one of the main reasons why the current LORs are still implemented over RDBMS systems [5]. This perception has been formed in the early days of RDF stores, with current systems promising improved performance. However, the perception persists. Moreover, to the authors knowledge, there are no studies focused on evaluating their performance in the context of e-learning systems. The present study stress tests the performance of two most successful RDF native database implementations that are openly available (4store and Jena Apache) in the specific context of a Semantic Learning Repository. It is known that there are already benchmarks proposed for the comparison of performance among different RDF database implementations such as Berlin SPARQL Benchmark (BSBM) [6] and Lehigh University Benchmark [7], as well as other studies involving performance experiments on RDF databases [8] [9], however, as stated by [10], there is a need for testing and comparing performances for each type of application in specifics architectures, contexts and scenarios, in this case in the common queries produced by an e-learning solution.

The reminder of this paper is organized as follows. Section 2 describes the materials and the methodology of the present study, and section 3 presents the results of the tests we performed. The final remarks are presented in section 4.

II. Materials and Methods

In order to determine which RDF database performs better, a set of queries specific to the learning context was created, executed and tested on identical conditions for both implementations: 4store and Jena Apache. The tests were performed for the data model presented in Fig. 1. This data model was implemented for the APRENDE Tutoring System.

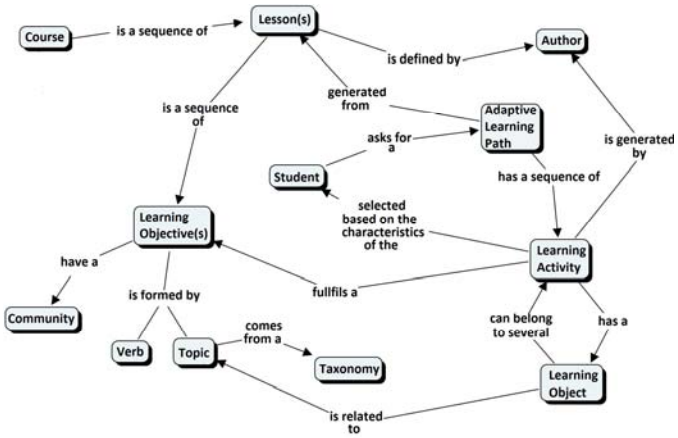

Fig. 1. Data Model

In APRENDE[1] teachers can upload learning materials (also denominated as learning activities) and associated them to learning objectives. Moreover, lessons (sequences of learning objectives) and courses (sequences of lessons) can be created and offered to the students. As the students navigate through the available lessons and courses, the system learns about their profile and recommends personalized learning materials to those lessons they are studying. A more detailed description of the APRENDE can be seen in [11]. The dataset contains a total of 1,262,954 triples, distributed over 16 graphs that were used for both evaluations. The proportional distribution of triples per graphs is shown in Table I. In order to avoid (reduce) the influence of network latency, both RDF implementations and the repository's website were running in the same machine. The hardware and software configuration are shown in Table II.

TABLE I
PERCENTAGES AND NUMBER OF TRIPLES PER GRAPHS

Graph	Percentage of total triples (%)	Number of triples
G1. Learning Activity Weight per User	80,55	1.017.339
G2. Learning Path Activity	15,29	193.140
G3. Learning Path	1,32	16.731
G4. Adaptation	0,71	9.011
G5. User	0,65	8.170
G6. Learning Style - User Profile	0,57	7.177
G7. Learning Activity	0,56	7.126
G8. Objective	0,15	1.869
G9. Learning Activity Objectives	0,06	750
G10. Lesson Objectives	0,05	663
G11. Lesson	0,04	545
G12. Course Lessons	0,02	231
G13. Course	0,008	101
G14. Taxonomy	0,005	60
G15. Sequence	0	6

TABLE II
HARDWARE AND SOFTWARE SETUP CONFIGURATION

Type	Component	Hardware/Software in use
Hardware	Processor	Intel(R) Core(TM) i5-2430M 2.40GHz 64bits (2 cores, 4 threads).
	Memory	8GB.
	Hard disk	250GB

[1] http://aprende.igualproject.org

Software	RDF Database	Apache Jena-Fuseki 1.1.1 with built-in TDB and Garlik 4store version 1.1.5.
	Operating System	Linux Fedora 19 64 bits.
	Filesystem	ext4.
	Java Version and JVM	Version 1.7.0_65, OpenJDK 64-Bit Server VM (build 24.65-b04)

The study focused on measuring two different aspects: 1) the resulting time to upload the triples in the databases; and 2) the resulting time to execute queries.

A. Implementation of the Databases

The implementation of the databases and the setup of the whole evaluation system consisted on the following steps: 1) Exporting RDF data that were already created to Turtle format[2]; 2) Initializing and executing each database; 3) Uploading the RDF data to the database currently under test; 4) Integrating the system website with the databases in order to allow RDF querying; and 5) Running the queries. The study focused on measuring the performance of steps 3 and 5 (upload and response time).

B. SPARQL queries

As the repository website was in PHP, we used Zend framework[3] to integrate it with the database in order to allow RDF querying. The Zend framework provides a HTTP Client class that allows one to send HTTP requests to an endpoint. This approach is possible because both 4store and Apache Jena provide RESTful API to the developer. In other words, there is a particular endpoint for the dataset and for each operation to be done on it (query, update, etc). Classic HTTP requests (POST, GET, PUT, HEAD, etc) can be sent to these endpoints to perform queries and updates. For the present study we tested queries and updates by using the POST method.

Apache Jena provides two alternatives for the upload of the datasets. The first one is the upload via a browser's interface at the regular address http://localhost:3030 and using a POST method to add triples in a previously created named graph. The second one is SOH (SPARQL Over HTTP), which is a set of Ruby scripts provided together with the Fuseki's distribution. In this case, the s-put script uses a PUT method to add triples in a named graph that may or not already exist. Both alternatives provided by Apache Jena were evaluated here.

The set of queries for evaluation was designed to be similar to the set that is executed when a given user access the repository. A total of 27 queries were evaluated. The queries can be divided in 4 major groups that are presented in Table IV. For the present study, the queries were executed in the same order that they are presented in Table III. This chronological sequence expressed by the execution of M1 + M2 + M3 + M4 is denominated here as Session and implies in a natural and realistic caching improvement for both databases. In other words, the Session is composed by a query mix that is naturally executed when a given user navigates on the system (the sequence of logging into the system, accessing a course, continuing a course, and completing a course). In our tests, all queries obey this chronological sequence represented by the Session. Differently from a benchmark work where a query is uninterruptedly executed numerous times, here all executions are performed in the sequence provided by the Session. Queries are complicated from M1 to M4. For instance, in M1, queries search results by using one or more graphs to authenticate users and show him their lessons and courses, whereas in M3, searches and new data insertions are executed over only one graph at a time.

During the tests, the Session was executed two times for each RDF implementation. For each Session execution, some queries of M1 and M2 groups were executed two times, in order to simulate a natural behavior of a user that visualizes a course and returns to the homepage twice. The average time of these two Session executions is then computed as the resulting time response for that given RDF implementation.

TABLE III
QUERIES

Group Mix	Queries Description	Graphs involved
M1 - Access to the site: logon process e homepage loading.	Use of FILTER, REGEX, ORDER BY, LIMIT, queries with and without FROM clause, queries with single and multiple variables, one graph search and multiple graphs search.	G1, G3, G5, G9, G10, G12, G13
M2 - Access to a course: course's content visualization, the start of a new course and creation of a new learning path.	Use of DISTINCT, ORDER BY, GROUP BY, COUNT, queries with and without FROM clause, one graph search, multiple graphs search and INSERT DATA queries	G1,G2, G3, G8, G9, G10, G12, G13, G15
M3 - Continuation of a course: learning path's information, selection and information of the next available activity, information of the current objective and activity, information of a lesson and a course's objective and completion of an activity	Use of FILTER, REGEX, LIMIT, ORDER BY, queries with single and multiple variables, one graph search, WITH DELETE INSERT queries.	G2, G3, G7, G8, G11 G13
M4 - Completion of a course: completion of the learning path.	WITH DELETE INSERT queries	G3

III. RESULTS

As mentioned earlier, the present study focused on the time response for uploading the RDF data to the databases, and for running the queries.

A. Uploading the RDF

Figure 2 shows the corresponding time in seconds to upload triples in the graphs in each database. In all cases 4store has better uploading time compared to the other two alternatives of Jena, but a significant difference is observed

[2] http://www.w3.org/TR/turtle/
[3] http://framework.zend.com/

when loading data in graphs with the largest number of triples (> 190000). For the graph G1 (1,017,339 triples), 4store took 8.90s, whereas Apache Jena with POST took 61.93s and Apache Jena with PUT took 245.13s.

B. *Running queries*

Table IV shows the total time for each mix of the Session in both implementations. As can be seen in Table IV, the time response of 4store is better than Apache Jena for all the scenarios. For instance, for the first and second mixes of the Session (M1 and M2), 4store responds 3.7 and 4.1 times faster than Apache Jena. When the user is in the middle of the course (M3), the performance of 4store is 5.7 times better than the Apache Jena. The only case where the time response for the queries in both implementations is similar is for M4 (WITH DELETE INSERT queries).

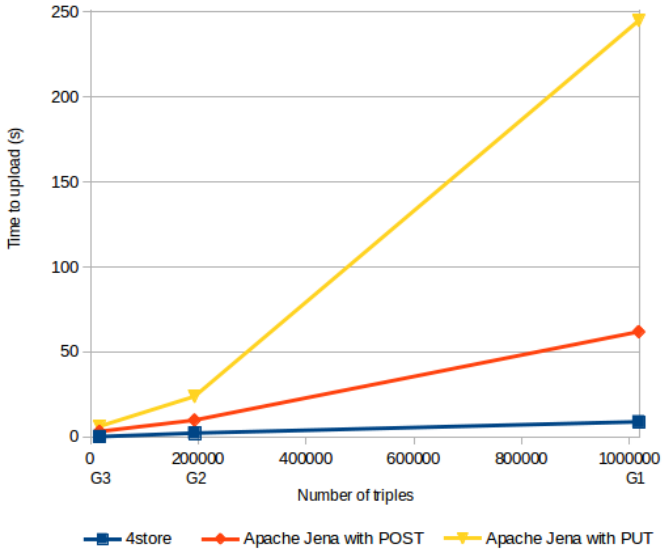

Fig. 2. Upload performance of the databases

TABLE IV
TIME RESPONSE FOR THE QUERIES

Queries Groups	Total types of queries	Total quantity of queries	Total execution time in 4store (ms)	Total execution time in Apache Jena (ms)
M1	5	16	178.66	660.20
M2	7	23	3,273.86	13,409.92
M3	10	11	253.53	1,444.36
M4	1	2	231.72	306.25

IV. FINAL REMARKS

Current Learning Object Repositories (LOR) have serious shortcomings that reduce their usefulness to implement e-learning systems that go beyond retrieving Learning Objects. The whole e-learning ecosystem demands a drastic change in the way we see, use and therefore, implement such repositories. The next LOR generation needs to consider and represent all the other entities that participate in the e-learning process (teachers, students, lessons, courses, activities, learning paths, etc.) in a way that they are all fully integrated and linked-up. We propose the use of Semantic Learning Repositories in order to provide the functionality needed by modern e-learning solutions.

This study evaluated the performance of two different RDF implementations (4store and Jena Apache) in the specific context of a Semantic Learning Repository called APRENDE. The results showed that 4store performed better than Apache Jena for all the scenarios we evaluated. More importantly, the times found for both solutions, especially for the one implemented with 4store are enough to implement an on-line system that could provide the required services to e-learning systems. Although this work is limited in scope, it is a first empirical proof that RDF stores could be successfully used to select Semantic Learning Repositories to implement larger learning solutions. Future work will focus on testing other RDF store implementations and evaluate other aspects of the system performance.

REFERENCES

[1] X. Ochoa and E. Duval, "Use of contextualized attention metadata for ranking and recommending learning objects," in *Proceedings of the 1st international workshop on Contextualized attention metadata: collecting, managing and exploiting of rich usage information*, 2006, pp. 9-16.

[2] B. Simon, D. Massart, F. Van Assche, S. Ternier, E. Duval, S. Brantner, *et al.*, "A simple query interface for interoperable learning repositories," in *Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems*, 2005, pp. 11-18.

[3] E. Bogdanov, C. Ullrich, E. Isaksson, M. Palmer, and D. Gillet, "From LMS to PLE: a step forward through opensocial apps in moodle," in *Advances in Web-Based Learning-ICWL 2012*, ed: Springer, 2012, pp. 69-78.

[4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International journal on semantic web and information systems*, vol. 5, pp. 1-22, 2009.

[5] M. Bergman, "Advantages and Myths of RDF," *AI3, April*, 2009.

[6] B. Christian and S. Andreas, "The Berlin SPARQL Benchmark," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 5, pp. 1-24, 2009.

[7] Y. Guo, Z. Pan, and J. Heflin, "LUBM: A benchmark for OWL knowledge base systems," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 3, pp. 158-182, 10// 2005.

[8] M. Schmidt, T. Hornung, N. Küchlin, G. Lausen, and C. Pinkel, "An Experimental Comparison of RDF Data Management Approaches in a SPARQL Benchmark Scenario," in *The Semantic Web - ISWC 2008*. vol. 5318, A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 82-97.

[9] C. Bizer and A. Schultz, "Benchmarking the performance of storage systems that expose SPARQL endpoints," *World Wide Web Internet And Web Information Systems*, 2008.

[10] Y. Guo, A. Qasem, Z. Pan, and J. Heflin, "A requirements driven framework for benchmarking semantic web knowledge base systems," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 19, pp. 297-309, 2007.

[11] X. Ochoa, G. Carrillo, and C. Cechinel, "Use of a Semantic Learning Repository to Facilitate the Creation of Modern e-Learning Systems.," in *Workshop EER: E-Learning and Educational Resources. XV International Conference on Human Computer Interaction (Interacción 2014)*, Puerto de la Cruz, Tenerife, 2014, pp. 535-542.