The SEALS Yardsticks for Ontology Management

Raúl García-Castro[1], Stephan Grimm[2], Ioan Toma[3],
Michael Schneider[2], Adrian Marte[3]

[1] Ontology Engineering Group, Departamento de Inteligencia Artificial.
Facultad de Informática, Universidad Politécnica de Madrid, Spain
rgarcia@fi.upm.es
[2] FZI Research Center for Information Technology, Karlsruhe, Germany
{grimm,schneid}@fzi.de
[3] STI Innsbruck. Universität Innsbruck, Austria
{ioan.toma,adrian.marte}@sti2.at

Abstract. This paper describes the first SEALS evaluation campaign over ontology engineering tools (i.e., the SEALS Yardsticks for Ontology Management). It presents the different evaluation scenarios defined to evaluate the conformance, interoperability and scalability of these tools, and the test data used in these scenarios.

1 Introduction

Ontology engineering tools are a cornerstone in the development of the Semantic Web and still they lack some set of common evaluations and test data that can be used to assess whether these tools are suitable for specific use cases.

In SEALS we aim to automatically evaluate ontology engineering tools. This implies a mayor challenge in this type of tools; first, because of the high heterogeneity between these tools and, second, because the perception over these tools is usually related to their user interfaces. Nevertheless, we plan to disregard user interaction from the evaluations (either from real users or simulated) and to measure the relevant characteristics of these tools through programmatic interactions in order to obtain fully-automatic evaluations.

The SEALS Yardsticks for Ontology Management is an evaluation campaign over ontology engineering tools that contains three evaluation scenarios for evaluating the conformance, interoperability and scalability of these tools, and that are supported by different evaluation services provided by the SEALS Platform.

The first characteristic that we will cover in the evaluation campaign will be the conformance of ontology development tools. Previously, this conformance has only be measured in qualitative evaluations that were based on tool specifications or documentation, but not on running the tools and obtaining results about their real behaviour. Some previous evaluations provided some information about the conformance of the tools since such conformance affected the evaluation results; however, the current situation is that the real conformance of existing tools is unknown. Therefore, we will evaluate the conformance of ontology engineering tools and we will cover the RDF(S) and OWL W3C recommendations.

A second characteristic that we will cover, highly related to conformance, is interoperability. Previously, in the RDF(S) and OWL Interoperability Benchmarking activities [1] the interoperability of several semantic technologies was evaluated using RDF(S) and OWL Lite as interchange languages. In this evaluation campaign we will extend these evaluations with test data for OWL DL and OWL Full to fully cover the RDF(S) and OWL recommendations.

Scalability is a main concern for any semantic technology, including ontology engineering tools. Nevertheless, only one effort was previously performed for evaluating the scalability of this kind of tools and it was specific to a single tool [2]. In this first evaluation campaign we will establish the grounds for the automatic evaluation of the scalability of ontology engineering tools, with the aim of proposing an extensible approach to be further extended in the future.

In all these evaluation scenarios, the only requirement for performing the evaluation on a tool is that the tool is able of importing and exporting ontologies in the ontology language. Therefore, the evaluations can be performed not only on ontology engineering tools but also on other types of semantic technologies.

2 Evaluation Scenarios

In the evaluation scenarios that compose the evaluation campaign, we need an automatic and uniform way of accessing most of the semantic tools and the operations performed to access such tools must be supported by most of them. Due to the high heterogeneity in semantic tools, ontology management APIs vary from one tool to another. Therefore, the way chosen to automatically access the tools is through the following two operations commonly supported by most semantic tools: to import an ontology from a file (i.e., to load an ontology from a file into the tool internal model), and to export an ontology to a file (i.e., to store an ontology from the tool internal model into a file).

The next sections describe the three evaluation scenarios used in the evaluation campaign. A detailed description can be found in [3].

Conformance. The conformance evaluation has the goal of evaluating the conformance of semantic technologies with regards to ontology representation languages, that is, to evaluate up to what extent semantic technologies adhere to the specification of ontology representation languages.

During the evaluation, a common group of tests is executed and each test describes one input ontology that has to be imported by the tool and then exported. After a test execution, we have two ontologies in the ontology representation language, namely, the original ontology and the final ontology exported by the tool. By comparing these ontologies we can know up to what extent the tool conforms to the ontology language.

Interoperability. The interoperability evaluation has the goal of evaluating the interoperability of semantic technologies in terms of the ability that such technologies have to interchange ontologies and use them. In concrete terms, the evaluation takes into account the case of interoperability using an interchange language.

During the experiment, a common group of tests is executed and each test describes one input ontology that has to be interchanged between a single tool and the others (including the tool itself). After a test execution, we have three ontologies in the ontology representation language, namely, the original ontology, the intermediate ontology exported by the first tool and the final ontology exported by the second tool. By comparing these ontologies we can know up to what extent the tools are interoperable.

Scalability. The scalability evaluation has the goal of evaluating the scalability of semantic technologies in terms of time characteristics. More concretely in our case, the scalability evaluation is concerned with evaluating the ability of ontology engineering tool to handle large ontologies.

During the evaluation, a common group of tests is executed and each test describes one input ontology that has to be imported by the tool and then exported. We are interested in the amount of time it takes to perform import and export operations on large size ontologies. After a test execution, we have as result two ontologies, the original ontology and the final ontology exported by the tool, and execution information including the time when the import and export operations started and ended.

3 Test Data

Conformance and Interoperability. In the first evaluation campaign, the conformance and interoperability evaluations will cover the RDF(S) and OWL specifications. To this end, we will use four different test suites that contain synthetic ontologies with simple combinations of components of the RDF(S), OWL Lite, OWL DL, and OWL Full knowledge models.

The RDF(S) and OWL Lite Import Test Suites already exist and detailed descriptions of them can be found in [1]. The OWL DL and OWL Full Import Test Suites have been developed in the context of the SEALS project and are described in [4].

Scalability. Two test suites were defined to be used for the scalability evaluations. The Real-World Ontologies Scalability Test Suite includes real-world ontologies in OWL DL that have been identified as being relevant for scalability evaluation [3]: AEO, the NCI Thesaurus, GALEN, the Foundational Model of Anatomy Ontology (FMA), the OBO Foundry, Robert's family ontology, and the wine and food ontologies. From this large set of ontologies we have selected 20 ontologies of various sizes to construct a first scalability test data suite.

The second test suite was defined using the Lehigh University Benchmark (LUBM)[1] data generator (UBA) that generates data over the Univ-Bench ontology[2], which describes universities and departments and the activities that occur at them [5].

4 Conclusions

Evaluation automation will allow to convert evaluations of ontology engineering environments from one-time evaluation activities to effortless continuous evaluations. Furthermore this reduction of effort in evaluating these tools will allow not only to perform evaluations with large test data, but also to perform evaluations that would be difficult or impossible to be performed manually. Equally important in the research area is the benefit of repeatability that automatic evaluations provide, allowing to perform evaluations multiple times in a consistent way and to objectively compare research findings.

All the resources used in this evaluation campaign as well as the results obtained will be publicly available through the SEALS Platform. This way anyone interested in evaluating an ontology engineering tool will be able to do so, and to compare to others, with a small effort.

Our future plans are to extend, on the one hand, the evaluation scenarios to cover more tool characteristics and, on the other hand, the evaluation data to include new test suites to cover the OWL 2 specification. With these extensions, we plan to conduct a second edition of this evaluation campaign.

Acknowledgements

This work has been supported by the SEALS European project (FP7-238975).

References

1. García-Castro, R.: Benchmarking Semantic Web technology. Volume 3 of Studies on the Semantic Web. AKA Verlag – IOS Press (2010)
2. García-Castro, R., Gómez-Pérez, A.: Guidelines for benchmarking the performance of ontology management APIs. In: Proceedings of the 4th International Semantic Web Conference (ISWC2005), Galway, Ireland, Springer (2005) 277–292
3. García-Castro, R., Grimm, S., Schneider, M., Kerrigan, M., Stoilos, G.: D10.1. Evaluation design and collection of test data for ontology engineering tools. Technical report, SEALS Project (2009)
4. García-Castro, R., Toma, I., Marte, A., Schneider, M., Bock, J., Grimm, S.: D10.2. Services for the automatic evaluation of ontology engineering tools v1. Technical report, SEALS Project (2010)
5. Guo, Y., Pan, Z., Heflin, J.: LUBM: A benchmark for OWL knowledge base systems. J. Web Sem. **3** (2005) 158–182

[1] `http://swat.cse.lehigh.edu/projects/lubm/`

[2] `http://www.lehigh.edu/~zhp2/2004/0401/univ-bench.owl`