Semantic Evaluation At Large Scale (SEALS)

Stuart N. Wrigley
Dept. Computer Science
University of Sheffield, UK
s.wrigley@dcs.shef.ac.uk

Raúl García-Castro
Facultad de Informática
Universidad Politécnica de Madrid, Spain
rgarcia@fi.upm.es

Lyndon Nixon
STI International
Vienna, Austria
lyndon.nixon@sti2.org

ABSTRACT

This paper describes the main goals and outcomes of the EU-funded Framework 7 project entitled Semantic Evaluation at Large Scale (SEALS). The growth and success of the Semantic Web is built upon a wide range of Semantic technologies from ontology engineering tools through to semantic web service discovery and semantic search. The evaluation of such technologies – and, indeed, assessments of their mutual compatibility – is critical for their sustained improvement and adoption. The SEALS project is creating an open and sustainable platform on which all aspects of an evaluation can be hosted and executed and has been designed to accommodate most technology types. It is envisaged that the platform will become the de facto repository of test datasets and will allow anyone to organise, execute and store the results of technology evaluations free of charge and without corporate bias. The demonstration will show how individual tools can be prepared for evaluation, uploaded to the platform, evaluated according to some criteria and the subsequent results viewed. In addition, the demonstration will show the flexibility and power of the SEALS Platform for evaluation organisers by highlighting some of the key technologies used.

Categories and Subject Descriptors

D.2 [**Software Engineering**]: Software/Program Verification; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation, Measurement, Performance

Keywords

Semantic Technology Evaluation, Semantic Search, Ontology Engineering, Ontology Matching, Semantic Web Services, Semantic Storage and Reasoning, Evaluation Infrastructure

1. INTRODUCTION

Semantic technologies play a critical role in the recent advances in both the Web (the *Semantic Web*) and corporate knowledge management. Such developments are revolutionising the way information and knowledge are processed. Semantic technologies provide ways to express knowledge and data so that they can be properly exploited by computers in an automated way for different purposes such as information retrieval or data integration.

The evaluation of such technologies is crucial for their sustained improvement and adoption, allowing users to assess the suitability of current technologies to their needs.

Some initiatives have already created a basis for semantic technology evaluation, such as those in the areas of ontology matching [5], ontology engineering [8, 9], ontology reasoning [12, 15], semantic search [13] or semantic web services [14, 17]. However, additional effort is required to accommodate the growth of the field, since evaluation is still costly, both in terms of reusing evaluation resources defined by others and of actually executing evaluations and analysing their results.

One clear direction for facilitating semantic technologies evaluation is the automation of evaluation processes. However, such automation is a complex task that requires: 1. the coordinated interaction in an evaluation workflow of all the involved resources, e.g., tools, test data and evaluation results; 2. the definition of such evaluation workflows in some machine-processable format; and 3. the ability to cope with the heterogeneity of the different tools and resources. Therefore, we have devised a solution for automated evaluation, within the context of the SEALS Project[1].

At the heart of the EU-funded Framework 7 SEALS Project is the development of the SEALS Platform [7]: an open infrastructure for the evaluation of semantic technologies that offers independent computational and data resources for the evaluation of those technologies. To this end, the SEALS Platform provides a common evaluation framework, based on the reusability of evaluation resources, in which different types of semantic technologies can be automatically evaluated. Indeed, the versatility of the platform was demonstrated during the first worldwide SEALS evaluation campaign held during mid-2010 in which tools from five different semantic technology fields (ontology engineering, semantic search, semantic web services, ontology matching, storage and reasoning) were formally evaluated [16].

In addition to large, formal evaluation campaigns, the SEALS Platform has also been designed to facilitate ad-hoc evaluations by individuals or organisations. To this end, use of the SEALS Platform and associated technologies is free of charge and all code is Open Source (Apache License v2.0).

WWW2012, Lyon, France
ACM 978-1-4503-1229-5/12/04.

[1] http://www.seals-project.eu/

2. TARGET TECHNOLOGIES

The SEALS Project has identified five core technology areas which lie at the heart of the Semantic Web. As such, these have been used to demonstrate the effectiveness and utility of the SEALS Platform. Not only do evaluations within these areas provide valuable case studies and proof-of-concept but they also provide invaluable insights into the technologies themselves; insights which can be, and are being, used to improve performance of Semantic Web tools.

2.1 Ontology Engineering Tools

Two types of tools support ontology engineering tasks: ontology editors, which are user-oriented and allow creating and maintaining ontologies mainly through user interfaces, and ontology management programming interfaces, which are developer-oriented and allow the creation and maintenance of ontologies through programming interfaces.

Since there exist different ontology languages (e.g., RDF-S, OWL, OWL 2), each with different expressiveness and reasoning capabilities, the conformance and interoperability of semantic technologies with regards to ontology language specifications is one of the main characteristics to evaluate in these tools. Conformance and interoperability evaluations [8, 9] use groups of ontologies defined in specific ontology languages as test data; these evaluations are performed by making tools process ontologies (coming either from test data or from other tools) and analysing the processed ontology (usually by comparing the processed ontology with that used as input).

2.2 Ontology Reasoning Tools

Description Logics (DLs) [2] are a family of logic-based knowledge representation formalisms designed to represent and reason about the knowledge of an application domain in a structured and well-understood way. Besides their formal knowledge representation languages, DLs also provide inference services. The aim of such services is to extract new implied information out of the explicitly stated information. Every knowledge representation language usually offers a different set of inference services. The most widely used inference services include: class satisfiability, classification, logical entailment, and ontology satisfiability.

In order to interact with other systems an ontology reasoner must conform to standard input formats and must be able to provide standard inference services. The performance criterion relates to an ontology reasoner's ability to efficiently perform these standard inference services.

2.3 Ontology Matching Tools

Matching ontologies consists of finding a set of correspondences (alignment) between two different ontologies. A wide diversity of systems have been proposed, which can be classified according to the many features that can be found in ontologies (e.g., labels, structures, instances, semantics), or with regards to the techniques they use (e.g., statistics, combinatorics, semantics, linguistics, or machine learning) [6].

The most commonly used criterion for evaluating matching systems is the compliance of matcher alignments with respect to the expected reference alignments. Metrics such as precision and recall are largely adopted for quantitatively evaluating matching tools. Other evaluation criteria are efficiency, in terms of runtime and memory consumption, and scalability using large sets of tests; semantic measures, where the proximity between alignments is measured instead of their strict equality [3, 4]; and task-specific evaluations, where alignments are evaluated according to their usage in some specific task.

2.4 Semantic Search Tools

State-of-the-art semantic search approaches are characterised by their high level of diversity both in their features as well as their capabilities. Such approaches employ different styles for accepting the user query (e.g., forms, graphs, keywords) [18] and apply a range of different strategies during processing and execution of the queries. They also differ in the format and content of the results presented to the user. All of these factors influence the user's perception of performance and usability.

Semantic search technologies can be evaluated on the basis of different criteria and metrics [19, 13]. At the core of any search task is the retrieval of pertinent information; search evaluations employ several questions which are applied to a particular ontology and dataset. Since (for ontology-based search) the answer set for each question is finite and known *a priori*, the measures of precision and recall are used. We are also interested in how tools cope with increasingly large datasets (scalability). Since search is an inherently user-oriented task, evaluation must also consider metrics such as how long it takes for a query to be executed.

2.5 Semantic Web Service Tools

Semantic Web Service (SWS) technologies enable the automation of discovery, selection, composition, mediation and execution of web services by means of semantic descriptions of their interfaces, capabilities and non-functional properties. SWS provide a layer of semantics for service interoperability by relying on a number of reference service ontologies and semantic annotation extension mechanisms.

The evaluation of SWS technologies is currently being pursued by a number of initiatives using different evaluation methods (e.g., see [14, 17]). Although these initiatives have succeeded in creating an initial evaluation community in this area, they have been hindered by the difficulties in creating large-scale test suites and by the complexity of manual testing to be done.

The SEALS Platform provides the infrastructure to homogenise these approaches and eliminate the necessity for time-consuming manual evaluation.

3. PROJECT OUTPUTS

There are three major outputs from the SEALS Project: the evaluation infrastructure (the SEALS Platform); the organisation and execution of two worldwide evaluation campaigns; and the creation / enhancement of a community of interest surrounding semantic evaluation and, more specifically, the SEALS technologies which can facilitate this.

3.1 SEALS Platform

The SEALS Platform is an open infrastructure for the evaluation of semantic technologies that offers independent computational and data resources for the evaluation of those technologies. To this end, the SEALS Platform provides a common evaluation framework, based on the reusability of evaluation resources, in which different types of semantic technologies can be automatically evaluated. It is responsible for all aspects of the evaluation: test data management;

tool configuration and execution; result generation and storage, etc. In order to ensure reproducibility and allow direct performance comparison, an entire evaluation is conducted within the SEALS Platform. In other words, all test data is stored locally as are the tools to be evaluated. The tools themselves are executed *within* the SEALS Platform (using virtual machine approaches to handle operating system dependencies) and once one or more tools have been evaluated, the generated results and any subsequent analyses are also stored locally and are made available for visualisation.

In addition to the core hardware, additional software components are in development to allow the SEALS Platform to be executed in cloud computing resources such as the Amazon Elastic Compute Cloud (Amazon EC2) facility.

The design of the SEALS Platform allows the SEALS framework to evaluate a variety of heterogeneous tool technologies, from different semantic areas, and is extendible to encompass new evaluations (i.e., different types of tools). Naturally, the steps necessary to evaluate an ontology matching tool will be very different from those to evaluate a semantic search tool, for instance. Therefore, we use the Business Process Execution Language (BPEL) [1] to provide an efficient means of scripting the entire lifecycle of a particular evaluation (a workflow). Furthermore, the adoption of an industry standard scripting approach (BPEL) facilitates SEALS Platform use by reducing conceptual overheads.

3.2 SEALS Evaluation Campaigns

The SEALS Platform is being used in two public worldwide evaluation campaigns and the results of these evaluation campaigns will be employed in creating semantic technology roadmaps, identifying sets of efficient and compatible tools for developing large-scale semantic applications. It is important to emphasise that these evaluations (and indeed the SEALS Platform itself) are targeted at both the commercial developer / adopter market as well as academic researchers.

The first of these campaigns was conducted in the Summer of 2010; 31 tools, from developers in 10 different countries, were evaluated across the five technology areas and the results of the campaign were disseminated at the ISWC workshop IWEST[2]. Further analysis of the campaign findings have been published in a variety of conferences (e.g., [16]). Furthermore, to promote adoption within the commercial sector, business-oriented whitepapers have been produced which describe both the evaluation approach [10] as well as the outcomes of the first campaign [11].

The second campaign is currently being executed and the results will be publicly available in mid-2012.

3.3 SEALS Community and Sustainability

SEALS is establishing and diffusing best practices in evaluation throughout the whole semantic technology community. To achieve this, the SEALS consortium have organised a number of workshops and tutorials at the premier academic conferences and industry events in the field to disseminate our work. To aid this, we have established a large, and growing, *SEALS Community*[3] who have access to the latest developments and materials produced by SEALS. The 'home' for the SEALS Community is the SEALS Portal which provides information about the SEALS initiative and its activities to interested parties (to encourage individuals to join); presents summaries of SEALS evaluation campaign activities and results; and gives community members privileged access to community tools. The SEALS Portal also provides a number of mechanisms for keeping up to date with SEALS activities: news sections, blog entries, RSS feeds and Twitter feeds. In addition to this, the SEALS Portal also provides online access to the full range of SEALS repositories (test data, results, tools) allowing management and downloading of datasets and results.

In order to ensure the sustainability of the SEALS initiative beyond the funded period of the project, the SEALS Project Management Board is currently in the process of creating a working group under the auspices of STI International[4] which will provide a home for the organisation of future evaluation campaigns, fund raising and training. Furthermore, SEALS has also aligned itself closely with existing evaluation efforts in order to act as a facilitator in future evaluation campaigns (e.g., the Ontology Alignment Evaluation Initiative[5]).

4. THE DEMONSTRATION

The demonstration will provide an insight into the user experience of participating in a SEALS evaluation. Since an evaluation consists of a number of different stages, the demonstration will focus on a number of different aspects which, taken as a whole, represent the full evaluation lifecycle from the participants' point of view. In order to participate in any evaluation hosted on the SEALS Platform, the user must register and enrol in one or more evaluations. The demonstration will show how the SEALS Portal is used for user management. In addition to this, we will show how to the use SEALS Portal's interface to access the test datasets which have already been stored on the SEALS Platform. There are a wide range of datasets appropriate to each technology area described in Section 2 which can be browsed or searched and the full dataset subsequently downloaded.

Before a tool can be benchmarked, it must be wrapped (a simple Java interface to allow bi-directional communication between the tool and the SEALS Platform) and packaged in an appropriate manner. We will use real tools to demonstrate how this is achieved and provide advice on how to package the attendees' own tools. The next stage, once a tool has been packaged, is to upload the tool to the SEALS Platform and enrol it into one of the five existing evaluation campaigns organised by SEALS.

We will show how this tool is then evaluated on the SEALS Platform. Although this operation is normally executed in a batch and transparent manner (no input from the user and no 'console'-like output), it will be possible to demonstrate this stage in such a way as to show the interactions between the various parts of the SEALS Platform and be able to observe, in real-time, the progress of the evaluation.

The final stage of the demonstration will show how the results of the tool evaluation can be viewed on the SEALS Portal and downloaded for further offline analysis.

Importantly, the demonstration will be designed in such a way as to be flexible so as to address the needs of evaluation *organisers* as well as participants. At each stage, it will

[2]Proceedings: `http://CEUR-WS.org/Vol-666/`

[3]`http://www.seals-project.eu/join-the-community/`

[4]`http://www.sti2.org/`

[5]`http://oaei.ontologymatching.org/`

be possible to provide an insight into the internal operation of the evaluation process to highlight the power and flexibility of the SEALS Platform. For example, the design of the BPEL workflow and its associated technologies can be demonstrated.

5. CONCLUSIONS

SEALS is relevant to the entire semantic technology community, namely, researchers in semantic technologies, tool developers, and users. Participation enables the establishment and diffusion of best practices in evaluation within the whole community and provides access to a set of services that support the whole life-cycle of the evaluation of these technologies. Therefore, this community will be able to evaluate tools by reusing evaluations provided by the SEALS Platform, define their own evaluations and access content stored in the SEALS Platform (test data, results, etc.).

The SEALS Project answers an urgent need felt by the semantic technology community for evaluation of semantic technologies. SEALS is creating worldwide impact, leading to a faster maturation of semantic technologies and increasing the adoption of research results by industry.

SEALS will change the way in which semantic technology is evaluated. The infrastructure developed within SEALS provides yardsticks for both industry and academia when they evaluate their applications and/or innovations. Indirectly, this is expected to help accelerate innovation in all those fields in which evaluation mechanisms are provided, as has been the case with both TREC benchmarks in Information Retrieval and TPC benchmarks in database research.

6. ACKNOWLEDGMENTS

The authors are partially supported by the SEALS EU FP7 project (IST-2009-238975).

7. REFERENCES

[1] A. Alves, A. Arkin, S. Askary, B. Bloch, F. Curbera, Y. Goland, N. Kartha, Sterling, D. König, V. Mehta, S. Thatte, D. van der Rijn, P. Yendluri, and A. Yiu. Web Services Business Process Execution Language Version 2.0. OASIS Standard Committee, April 2007.

[2] F. Baader, D. McGuinness, D. Nardi, and P. Patel-Schneider. *The Description Logic Handbook: Theory, implementation and applications*. Cambridge University Press, 2002.

[3] M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In B. Ashpole, M. Ehrig, J. Euzenat, and H. Stuckenschmidt, editors, *Proc. of the Workshop on Integrating Ontologies*, volume 156, page 8. CEUR-WS.org, August, 2005 2005.

[4] J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. of the 20th International Joint Conference on Artificial Intelligence*, pages 248–253, Hyderabad, India, 2007.

[5] J. Euzenat, A. Ferrara, C. Meilicke, A. Nikolov, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Sváb-Zamazal, V. Svátek, and C. Trojahn dos Santos. Results of the Ontology Alignment Evaluation Initiative 2010. In *Proc. of the 5th Workshop on Ontology Matching*, pages 85–117, Shanghai, China, 2010.

[6] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

[7] R. García-Castro, M. Esteban-Gutiérrez, and A. Gómez-Pérez. Towards an infrastructure for the evaluation of semantic technologies. In *Proc. of the eChallenges 2010 Conference*, October 27-29 2010.

[8] R. García-Castro and A. Gómez-Pérez. RDF(S) Interoperability Results for Semantic Web Technologies. *International Journal of Software Engineering and Knowledge Engineering*, 19(8):1083–1108, 2009.

[9] R. García-Castro and A. Gómez-Pérez. Interoperability results for Semantic Web technologies using OWL as the interchange language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8:278–291, November 2010.

[10] R. García-Castro and S. N. Wrigley. SEALS methodology for evaluation campaigns. Technical report, SEALS Consortium, 2011.

[11] R. García-Castro, M. Yatskevich, C. T. Santos, S. N. Wrigley, L. Cabral, L. Nixon, and O. Zamazal. The state of semantic technology today – overview of the first SEALS evaluation campaigns. Technical report, SEALS Consortium, 2011.

[12] I. Horrocks and P. F. Patel-Schneider. DL systems comparison. In *Proc. of the 1998 Description Logic Workshop*, volume 11, pages 55–57, 1998.

[13] E. Kaufmann. *Talking to the Semantic Web — Natural Language Query Interfaces for Casual End-Users*. PhD thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, September 2007.

[14] M. Klusch, A. Leger, D. Martin, M. Paolucci, A. Bernstein, and U. Kuester. Annual International Contest S3 on Semantic Service Selection. `http://www-ags.dfki.uni-sb.de/~klusch/s3/`.

[15] F. Massacci and F. M. Donini. Design and results of TANCS-2000 non-classical (modal) systems comparison. In *Proc. of the International Conference on Automated Reasoning with Analytic Tableaux and Related Methods*, pages 52–56, London, UK, 2000. Springer-Verlag.

[16] L. Nixon, R. García-Castro, S. N. Wrigley, M. Yatskevich, C. T. D. Santos, and L. Cabral. The state of semantic technology today – overview of the first seals evaluation campaigns. In *Proc. of the 7th International Conference on Semantic Systems*, 2011.

[17] C. Petrie, T. Margaria, H. Lausen, and M. Zaremba. *Semantic Web Services Challenge: Results from the First Year*. Springer, 2009.

[18] V. Uren, Y. Lei, V. Lopez, H. Liu, E. Motta, and M. Giordanino. The usability of semantic search tools: a review. *The Knowledge Engineering Review*, 22(4):361–377, 2007.

[19] S. N. Wrigley, D. Reinhard, K. Elbedweihy, A. Bernstein, and F. Ciravegna. Methodology and campaign design for the evaluation of semantic search tools. In *Proc. of the International Workshop on Semantic Search*, 2010.