M. Brian Blake · Liliana Cabral
Birgitta König-Ries · Ulrich Küster
David Martin *Editors*

Semantic Web Services

Advancement through Evaluation

Springer

Semantic Web Services

M. Brian Blake • Liliana Cabral
Birgitta König-Ries • Ulrich Küster
David Martin
Editors

Semantic Web Services

Advancement through Evaluation

Springer

Editors
M. Brian Blake
University of Miami
Coral Gables, FL
USA

Liliana Cabral
The Open University
Knowledge Media Institute - KMI
Milton Keynes
UK

Birgitta König-Ries
University of Jena
Jena
Germany

Ulrich Küster
University of Jena
Jena
Germany

David Martin
Apple, Inc.
Cupertino, CA
USA

ISBN 978-3-642-28734-3 ISBN 978-3-642-28735-0 (eBook)
DOI 10.1007/978-3-642-28735-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012941919

ACM Codes: H.3, I.2, D.2, C.4

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

This book is about evaluating semantic web services. Obviously, this is a task for Heroes and most of us would rather clean up the Augean stables than perform this task. Nevertheless, it has been done, and I have been asked to provide my insight on it. As every child knows, big things should become understandable by breaking them into pieces. Obviously, this book is about the following elements: Evaluation, Semantics, Web, Services, Semantic Web, Web Services, Semantic Web Services, and Evaluating Semantic Web Services. Let's go through them step by step.

Evaluation is a tricky task. In the last millennium I attended a workshop on it (during a very hot summer in Budapest). There they evaluated heuristic search methods. In order to prevent any artificial bias, they used randomized data for it. At first, this sounded very reasonable, especially because the workshop chair had such an impressive and marvelous Oxford English accent – giving you the impression that you were actually speaking to Newton himself. Still, I was a bit surprised about the documented and mostly negative results. I started to wonder whether random data are the right resource to evaluate heuristics. Heuristics make certain assumptions about domain and task specific regularities in order to outperform generic search methods. Obviously the data used in this workshop prevented us from any bias, but is it not precisely a certain bias that makes heuristics work if the bias is chosen well? In other words, could you measure the added value of intelligence in a completely randomized (or alternatively completely frozen) universe? It was evolution that granted our bias the ability to survive in the environment with which we are confronted and which we continue to form according to this bias. In the end, I started to wonder whether not having a bias is actually a very powerful way of actually having one without being able to talk about it. From this experience I learned that you cannot escape your bias; your perception is focused, and something completely random has a rather limited bias and focus. You should rather make your bias explicit and an object of discourse. With this you do not escape it but you can partially observe and rationalize it. Quite an insight for a hot summer that smelted away objectivity as an illusion of people that seem to negate but actually absolute their subjectivity as a matter outside of any discourse!

Semantics is an even stranger beast. In [1] we made an effort to define it by using some others words in a structured, natural language sentence. Unfortunately, we ended up with a set of other words that were just as difficult to understand. Recursively expanding their natural language definitions brought us back in no more than seven steps from our point of departure. In the end, semantics seems to be defined through being semantics. In a certain sense, this should not really be a surprise. If you have a limited number of words to define their meanings, you quickly return to the word you are trying to define. This should not be a problem for most words, but slightly disappointing for words that are about meaning. Now what is the Origin of Meaning? When do I think you understood what I was talking about? When you perform in the way that I had hoped you should act. The meaning of the act of communication is rooted outside the sphere of communication, reflecting the fact that communication is just a partial aspect of structured cooperation process. Or, in the words of Bill Clinton: "It is the cooperation, stupid!" In the end, it is the usage of something that defines its meaning for the subject that is using it (and, by the way, also for the object).

Capturing the essence of the **Web** seems to be rather trivial compared to evaluating semantics. It was invented by Sir Tim just as he invented hypertext, the internet, computers, electricity, and gravitation. More seriously spoken, it was a more focused innovation and somehow a tiny step. He allowed pointers to point beyond the borderlines of existing hypertext systems and he used the internet protocol to implement these links. This was a small step for him but a significant step for mankind. He generated a new mass media on a global scale, with 404 as its bug and major feature. It is currently evolving from a web of documents into in a web of data and hopefully soon into a web of services, processes, sensors, streams, devices and many more. Some of us have become gray haired whilst waiting for this "Future Internet" that is more than just a large pile of static documents and data. When these gray haired people talk about the web they mean it as a synonym for large, decentralized, distributed, and heterogeneous networks no matter which specific protocol instance is used to implement them.

Services started as a verbal cover for a statistical anomaly. Economic activities that could neither be classified as primary nor secondary (agriculture and manufacturing) needed a label, especially because this exception slowly started to become the major economic activity in developed countries. Similar to the case of IBM when it gave up its traditional core business and needed a name and a vision to justify its future, covering with a slogan, a new and not very well understood area does not necessarily lead to good definitions of the field. According to Wikipedia, Services are the "soft parts of the economy"[1] and many of its characterizations read "soft", too, mostly only concrete in what services are not. I tend to understand services as a certain functionality that is provided in abstraction of the infrastructure that is providing it. In conclusion, when you are describing a service as a service you focus on what it is providing (its functionality) and not on how it is implemented.

[1] http://en.wikipedia.org/wiki/Tertiary_sector_of_the_economy

Therefore, services are not about tangible products but about an organized way to use these things as a means of achieving certain goals.

In contrast to their name **Web Services** do not have much to do with the web other than using XML as exchange syntax. However, they come with their own protocol (SOAP), and use a message-centric paradigm. Finally, most of them are not used on the web but in intranets neither being open available nor using a web protocol. Meanwhile, this is slightly chanced by a number of services being directly accessible on the web using HTTP as their protocol. However, this introduces a new difficulty. In the old days, one could argue that a web service is a URI described by a WSDL file. The new type of services usually does not have such a machine readable description. It is hard to distinguish an ordinary web site and a web service. Somehow this is not surprising since we inherit this difficulty from the vague definition of what a service is. Still, we can identify two major characteristics of services:

- They are means to encapsulate data. Take the multiplication of two digits as an example. Instead of materializing all possible results in a large and potentially infinite matrix one can publish a function that does these calculations when needed.
- They are means to perform transactions like buying a book or booking a journey.

The **Semantic Web** applies semantics to the web. Therefore, its first generation was document-centric. It provides annotations for describing web content. With the web of data, it evolved towards a means of directly providing data on the web without being structured as documents. Indeed, a SPARQL endpoint in the web of data could be viewed as a service, however, as a pure data delivery service. **Semantic Web Services** provide semantic annotations for web services. Since the field of web services is still in its infancy, semantic web services are nevertheless mostly an academic exercise compared to the huge take-up of the semantic web and the web of data. Also, they are tackling a much more difficult problem. They do not simply annotate a piece of data but a piece of software with potentially real world activities following their usage. Clearly, pragmatic assumptions must be made to save us from the impossibility of automatic programming.

Therefore, **Evaluating Semantic Web Services** is obviously a difficult task. A first step in this direction was made by Petrie et al. [2] and I congratulate the editors and authors of this issue for making a second one. It provides a complete and up-to-date survey of the field by integrating results from all major evaluation initiatives such as the Semantic Service Selection contest, the Semantic Web Service Challenge, and the Web Service Challenge. In conclusion, I can strongly recommend this book and it is a pleasure to provide a Foreword for it.

Innsbruck Dieter Fensel

References

1. J. Domingue, D. Fensel, J.A. Hendler, Introduction, in *Handbook of Semantic Web Technologies*, ed. by J. Domingue, D. Fensel, J.A. Hendler, vol. 1 (Springer, Heidelberg/New York, 2011)
2. C. Petrie, T. Margaria, H. Lausen, M. Zaremba (eds.), *Semantic Web Service Challenge* (Springer, New York, 2009)

Preface

This book compiles the perspectives, approaches and results of the research associated with three current Semantic Web Service (SWS) evaluation initiatives, namely the Semantic Service Selection (S3) contest,[1] the Semantic Web Service Challenge (SWS Challenge)[2] and the Web Service Challenge (WS Challenge).[3] The book will contain an overall overview and comparison of these initiatives as well as chapters contributed by authors that have taken part in one or more of these initiatives.

In addition, the participants are given the opportunity to focus on a comparative analysis of the features and performance of their tools with respect to other contest entries.

The goals of this book are to:

- Report results, experiences and lessons learned from diverse evaluation initiatives in the field of Semantic Web Services.
- Enable researchers to learn from and build upon existing work (SWS technology) and comparative results (SWS technology evaluation).
- Provide an overview of the state of the art with respect to implemented SWS technologies.
- Promote awareness among users and industrial tool providers about the variety of current Semantic service approaches.
- Provide information to enhance future evaluation methodologies and techniques in the field.

This book is aimed at two different types of readers. On the one hand, it is meant for researchers on SWS technology. These researchers will obtain an overview of existing approaches in SWS with a particular focus on how to evaluate SWS technology. In this community, the book will also encourage more thorough and

[1] http://dfki.uni-sb.de/~klusch/s3/index.html

[2] http://sws-challenge.org

[3] http://wschallenge.org/

methodological evaluation of new approaches. On the other hand, this book is meant for potential users of Semantic Web service technology and will provide them with an overview of existing approaches including their respective strengths and weaknesses and give them guidance on factors that should play a role in evaluation.

We hope the broader community will benefit from the insights gained from the experimental evaluation of the presented technologies. This book will extend the state of the art, which is concerned with developing novel technologies but often omits the experimental validation and explanation of their merits.

We would like to thank all the participants of the evaluation initiatives, who through their contributions promoted advances in the Semantic Web Service area.

The Editors (alphabetically):
M. Brian Blake
Liliana Cabral
Birgitta König-Ries
Ulrich Küster
David Martin

Contents

Chapter 1
Introduction

M. Brian Blake, Liliana Cabral, Birgitta König-Ries, Ulrich Küster, and David Martin

This introduction will provide the necessary background on Semantic Web Services and their evaluation. It will then introduce SWS evaluation goals, dimensions and criteria and compare the existing community efforts with respect to these. This allows comprehending the similarities and differences of these complementary efforts and the motivation of their design.

Finally, in the last section, we will discuss lessons learned that concern all of the evaluation initiatives. In addition, we will analyze open research problems in the area and provide an outlook on future work and directions of development.

1.1 Organization of the Book

The remainder of the book is divided into four parts. Each part refers to one of the evaluation initiatives, including an introductory chapter followed by chapters provided by selected participants in the initiatives.

M.B. Blake (✉)
University of Miami, Coral Gables, FL, USA
e-mail: M.Brian.Blake@miami.edu

L. Cabral
KMi, The Open University, Milton Keynes, UK
e-mail: l.s.cabral@open.ac.uk

B. König-Ries · U. Küster
Department of Computer Science, University Jena, Jena, Germany
e-mail: Birgitta.Koenig-Ries@uni-jena.de; Ulrich.Kuester@uni-jena.de

D. Martin
Apple, Inc., Cupertino, CA, USA
e-mail: david_martin@apple.com

M.B. Blake et al. (eds.), *Semantic Web Services*, DOI 10.1007/978-3-642-28735-0_1,

Part I will cover the long established first two tracks of the Semantic Service Selection (S3) Contest – the OWL-S matchmaker evaluation and the SAWSDL matchmaker evaluation. Part II will cover the new S3 Jena Geography Dataset (JGD) cross evaluation contest. Part III will cover the SWS Challenge. Finally, Part IV will cover the semantic aspects of the WS Challenge.

The introduction to each part provides an overview of the evaluation initiative and overall results for its latest evaluation workshops. The following chapters by the participants, in each part, will present their approaches, solutions and lessons learned.

1.2 SWS in a Nutshell

Semantic Web Services (SWS) has been a vigorous technology research area for about a decade now. As its name indicates, the field lies at the intersection of two important trends in the evolution of the World Wide Web. The first trend is the development of Web service technologies, whose long-term promise is to make the Web a place that supports shared activities (transactions, processes, formation of virtual organizations, etc.) as well as it supports shared information [3]. In the shorter term, the driving objective behind Web services has been that of reliable, vendor-neutral software interoperability, across platforms, networks, and organizations. A related objective has been the ability to coordinate business processes involving heterogeneous components (deployed as services), across ownership boundaries. These objectives, in turn, have led to the development of widely recognized Web service standards such as the Web Services Description Language (WSDL[1]) for specification of Web services and Universal Description, Discovery and Integration (UDDI[2]) for the advertisement and discovery of services. At least initially, Web services based on WSDL have been widely adopted in industry practices where interoperation could only be achieved through syntactic approaches. Inter-organization use of Web services, operated either by SOAP[3] or Representational State Transfer (REST) protocols, were limited to pre-established understanding of message types or shared data dictionaries.

Consequently, the second trend, the Semantic Web, is focused on the publication of more expressive metadata in a shared knowledge framework, enabling the deployment of software agents that can make intelligent use of Web resources or services [1]. In its essence, the Semantic Web brings knowledge representation languages and ontologies into the fabric of the Internet, providing a foundation for a variety of powerful new approaches to organizing, describing, searching for, and reasoning about both information and activities on the Web (or other networked environments).

[1] http://www.w3.org/TR/wsdl

[2] http://www.uddi.org/pubs/uddi_v3.htm

[3] http://www.w3.org/TR/soap

The central theme of SWS, then, is the use of richer, more declarative descriptions of the elements of dynamic distributed computation – services, processes, message-based conversations, transactions, etc. These descriptions, in turn, enable fuller, more flexible automation of service provision and use, and the construction of more powerful tools and methodologies for working with services.

Because a rich representation framework permits a more comprehensive specification of many different aspects of services, SWS can provide a solid foundation for a broad range of activities throughout the Web service lifecycle. For example, richer service descriptions can support greater automation of service discovery, selection and invocation, automated translation of message content (mediation) between heterogeneous interoperating services, automated or semi-automated approaches to service composition, and more comprehensive approaches to service monitoring and recovery from failure. Further down the road, richer semantics can help to provide fuller automation of such activities as verification, simulation, configuration, supply chain management, contracting, and negotiation of services. This applies not only to the Internet at large, but also within organizations and virtual organizations.

SWS research, as a distinct field, began in earnest in 2001. In that year, the initial release of OWL for Services (OWL-S, originally known as DAML-S) was made available[4] [14]. Other major initiatives began work not long thereafter, leading to a diverse array of approaches including the Web Services Modeling Ontology (WSMO[5], WSMO-Lite[6]), the Semantic Web Services Framework (SWSF[7]), MicroWSMO[8], and the Internet Reasoning Service (IRS [5]).

In the world of standards, a number of activities have reflected the strong interest in this work. Two of the most visible of these are Semantic Annotations for WSDL (SAWSDL[9]), which received Recommendation status at the World Wide Web Consortium (W3C) in August 2007, and SA-REST[10].

1.3 Evaluation in General

Evaluation has been part of science and scientific progress for a long time. In this section, we will have a brief look at evaluation in general before we focus on the much shorter history of evaluation in computer science.

[4] http://www.w3.org/Submission/OWL-S/

[5] http://www.wsmo.org/

[6] http://www.w3.org/Submission/2010/SUBM-WSMO-Lite-20100823/

[7] http://www.w3.org/Submission/SWSF/

[8] http://www.wsmo.org/TR/d38/v0.1/

[9] http://www.w3.org/2002/ws/sawsdl/

[10] http://www.w3.org/Submission/SA-REST/

1.3.1 Benefits and Aims of Evaluation

Lord Kelvin reportedly said more than 100 years ago, "If you can not measure it, you can not improve it". This sentence provides one of the main motivations for evaluations in a nutshell: By defining criteria that measure how good a system is, it becomes possible to objectively find strengths and weaknesses of this system and to systematically identify areas that need improvement. The German Evaluation Society puts it a bit more formally [2]:

> Evaluation is the systematic investigation of an evaluand's worth or merit. Evaluands include programs, studies, products, schemes, services, organizations, policies, technologies and research projects. The results, conclusions and recommendations shall derive from comprehensive, empirical qualitative and/or quantitative data.

When looking at the evaluation of software, [7] offers a useful summary of possible goals of an evaluation: It may aim at comparing different software systems ("Which one is better?"), at measuring the quality of a given system ("How good is it?") and/or at identifying weaknesses and areas for improvement ("Why is it bad?").

Despite it being obvious that asking the questions above makes sense and will contribute to advancing computer science, evaluation is – in general – rather neglected in computer science. While benchmarks etc. have long been used systematically in some areas of computer science, overall, systematic experimentation has only recently gained importance in other areas of computer science. This may be due to the fact that this is a very young discipline which didn't have much time yet to establish its scientific standards. Several independent studies show that compared to other sciences experimental papers and meaningful evaluations are less frequent in computer science [6, 16]. This hinders progress and makes adaptation of research results in industry more difficult since often no proven benefits exist [17]. An area of computer science where this has been recognized early on and has been overcome by a community effort, namely the establishment of the TREC conference, is Information Retrieval. This is particularly interesting in the context of this book, as Information Retrieval (IR) and Semantic Web Service Discovery have a number of obvious similarities (albeit also differences) that are leveraged by some of the initiatives described in this book. Many Semantic Web Service evaluation techniques duplicate and extend established IR quality measures.

1.3.2 Quality Criteria for Evaluation

Before delving into evaluation for Semantic Web Services, we will take a closer look at evaluation in general in this section. More particularly, we will review research on criteria that make an evaluation meaningful. Which criteria does an initiative need

to meet in order to come up with results that are useful and will really achieve the aims pursued by evaluations?

More systematically, the evaluation standards by the German Evaluation Society identify 25 requirements categorized in four groups that evaluations need to meet. Very briefly, these requirements are:

- Utility Requirements: Stakeholders should be identified and be able to become involved; the purpose of the evaluation needs to be explicitly identified; evaluators need to be trustworthy and competent; information needs of different parties need to be taken into consideration; evaluation results shall be timely reported in a complete and understandable manner.
- Feasibility Requirements: Evaluations shall be carried out in a cost-effective manner and in a way that maximizes acceptance by the stakeholders.
- Propriety Requirements: Obligations of the involved parties need to be made explicit; the rights of all stakeholders need to be preserved, the evaluation and reporting shall be fair, complete and unbiased;
- Accuracy Requirements: It should be explicitly described what is to be evaluated and in which context, what purposes and procedures are relevant and which information sources are being used; the collected data shall be systematically examined with respect to errors, qualitative and quantitative information; findings shall be justified; the evaluation itself should be evaluated.

It has been shown that community efforts are a good basis for meeting at least some of these criteria. The main advantages of community efforts are that they distribute the significant burden of evaluation and the development of appropriate test sets, criteria, measures, and so on, required by many participants. This is often the only feasible way to manage the overall burden and the most likely approach for the evaluation effort to be complete. Also, community efforts by their nature include many different aspects and view points and thus have a much better chance at being fair and unbiased than any effort by a single group or person. Additionally, community efforts offer a certain guarantee that all findings – and not only those convenient to a specific evaluator – will be reported on. Finally, by the involvement of a significant part of a research community in the evaluation initiative, a deeper understanding of the goals of a certain endeavor, appropriate means to quantify and measure achievement of these goals and of the area in general, will be widespread – and will further future progress in that area.

1.4 Evaluation for SWS

Now, that we have set the stage, let us have a closer look at SWS evaluation by first reexamining the aims of such an evaluation and then identifying dimensions for evaluation.

1.4.1 Aims of SWS Evaluation

Over the last decade, a vast amount of funding has been spent on the development of Semantic Web Service frameworks. Numerous description languages, matchmaking and composition algorithms have been proposed. Nevertheless, when faced with a real-world problem, it is, today, very hard to decide which of those different approaches to use. The situation was even worse 5 years ago, when there was basically the same plethora of approaches, but very few evaluations. To make things worse, these evaluations were done by different groups for their respective technologies without a common set of services or measures. So even where there existed evaluations, they could not be used to compare different approaches. This had at least two negative effects: First, it was (and to a degree still is) a major hurdle on the way to real-life adaptation of SWS technology. Potential users just did not know which technology was suitable for their problem – and they had no way of finding it out. Second, the lack of measurements and comparisons hindered the further advancement of science [13].

This situation is quite similar to the one observed by the IR community several decades ago:

> First [..] there has been no concerted effort by groups to work with the same data, use the same evaluation techniques, and generally compare results across systems. [...] The second missing element, which has become critical [..] is the lack of a realistically-sized test collection. [...] The overall goal of the Text REtrieval Conference (TREC) was to address these two missing elements. It is hoped that by providing a very large test collection, and encouraging interaction with other groups in a friendly evaluation forum, a new thrust in information retrieval will occur [8].

The enormous effect the concerted effort towards evaluation had on the IR community – but also similar effects observed in other communities creating benchmarks – is a strong incentive for similar efforts in SWS evaluation.

1.4.2 Dimensions of SWS Evaluation

Before we can start to evaluate, we need to decide on what we actually want to evaluate. In [12] a number of dimensions for evaluation of SWS, i.e., interesting aspects, were identified. For each of these aspects, an evaluation initiative will have to determine appropriate measurements and how they can be obtained. The dimensions are performance/scalability, usability/effort, correctness/automation, coupling, and functional scope/context assumptions, as described in the following.

Performance/Scalability This is probably the most obvious of the five dimensions and also the one where existing methods can probably be most easily adapted. It measures the resource consumption. Possible measures include runtime, memory consumption etc.

Usability/Effort The best solution does not help, if no one (or only few experts) can use it or if the effort of setting up the system is prohibitively high, thus, this effort should be measured by an evaluation. Quite obviously, it is not easy to find appropriate measures to capture this.

Correctness/Automation One of the most obvious criteria for a solution is that the results returned by a framework are correct. Here, IR-like measures including precision and recall (or variants thereof that take the subtle differences between IR and SWS into account) are used.

Coupling Here, criteria are needed that measure whether service offers and requests can be developed independently of another or not.

Functional Scope/Context Assumptions SWS frameworks differ widely in the functional scope they support: This ranges from static discovery over contracting and negotiation capabilities to automatic invocation and mediation.

1.5 Comparison of Current SWS Evaluation Initiatives

Now that we know the nature of the evaluations and what are the most meaningful criteria to regard with respect to SWS, let us briefly introduce the existing SWS evaluation initiatives and compare them according to the criteria and dimensions described in the previous sections. The initiatives will be described in detail in the introduction of the following parts.

For each of these initiatives, we will summarize their approach and give a short overview on the dimensions they address. We will also discuss how well they fit with the quality criteria for evaluation. The results of the comparison are summarized in Table 1.1 which has been adapted from [10]. In the latter, you'll find a much more detailed discussion of this topic.

1.5.1 The SWS Challenge

The SWS Challenge, originally founded by STI Innsbruck and the Stanford Logic Group, has been running as a series of workshops since 2006. As published in [15], its aim is to "develop a common understanding of various technologies and to explore the trade-offs among existing approaches".

The SWS Challenge provides a set of scenarios focusing on different aspects of the SWS problem space. Participants develop solutions to these scenarios and present these – including a code inspection – at the SWSC workshops. The scenarios fall in two broad categories namely mediation and discovery. For most of the scenarios, a testbed has been implemented. Solutions are supposed to be programmed against this testbed and need to actually call and execute the appropriate services. A lot of effort went into defining the evaluation methodology which was continuously adapted and refined over the course of the workshops.

Table 1.1 Existing initiatives in comparison

		SWS challenge	S3 contest	WS challenge
Dimension	Performance and scalability	n/a	Runtime for matchmaking	Runtime for composition
	Usability and effort	Adaptation effort	Description effort (cross eval track)	n/a
	Correctness and automation	No notion of partial correctness	Retrieval correctness	Correctness of algo, but not semantics
	Coupling	n/a	Decoupled setting, explicit in cross eval track	n/a
	Functional scope	Hierarchy of scenarios	Static discovery	Static composition
Criteria	Utility	++	++	+
	Feasibility	+	+	0
	Propriety	+	0	0
	Accuracy	−	0	+

In particular, measures were sought to capture the effort involved in adapting solutions to slight changes in the scenario. Initially, the idea was, that ideally, this should be possible without any programming effort by just changing declaration. However, it proved difficult to distinguish the two in practice.

The SWS Challenge concentrates on evaluating the functional scope of a framework. To a certain degree usability/effort are taken into consideration as well. Correctness/Automation are not measured; a proposed solution is either correct (i.e., provides the expected results) or not. There is no notion of a partially correct solution in the SWS Challenge. The other dimensions are not covered by the SWS challenge. Coupling has been paid no attention to at all (in general, offers and requests are written by the same people). Concerning performance/scalability this has not been an issue either. On the one hand, from a philosophical point of view the initiators of the SWSC did not deem them as important as other dimensions, on the other hand, the design of the SWSC is not suitable to measure performance. This is mainly due to the small number of scenarios which does not allow for statistically relevant performance measures.

Concerning the criteria for evaluations, the SWSC does pretty well with respect to utility, feasibility, and propriety requirements (with the exception of the need for formal agreement of stakeholders tasks). Its design results in less good marks concerning accuracy requirements. Since a lot of the evaluation is done in an interactive process at the workshops with manual code inspections and discussions, not all of the information is as "hard" as the accuracy requirements would like to see. Also, a meta-evaluation has been lacking, albeit that has been partially done by Küster [10].

1.5.2 The S3 Contest

The S3 (Semantic Service Selection) Contest was founded in 2006 by Mathias Klusch from DFKI Saarbrücken (Germany) and has been run annually together with groups from France Telekom Research, SRI International, NTT DoCoMo Research Europe, and the Universities of Zurich, Southhampton, and later on also Jena. It has an open call; results are presented annually at a workshop. The S3 contest performs evaluation in a number of tracks related to static discovery. These tracks either investigate the runtime performance and correctness of matchmakers in a single formalism or compare results across different formalisms. The S3 contest provides an extensive (albeit artificial) collection of semantically described services in different formalisms (OWL-S, SAWSDL) and a testing platform. Participants program their matchmaker against this platform. The platform will run the test and compute measures like run time, precision, recall and so on.

The S3 Contest uses an evaluation methodology that has long been agreed upon in the IR community. However, the adaptation to the SWS context raises some issues: First of all, the quality of the evaluation results depends strongly on the quality of the test collections. While OWLS-TC and SAWSDL-TC are no doubt the most comprehensive SWS test collections available and have been put together with considerable effort, there has been some concern about their quality. A more realistic collection would certainly be beneficial. Also, there exists a wide variety of measures for precision and recall or variants thereof. Up to now, a careful evaluation of which of these measures are best suited for SWS evaluation and what influence the measures have on the outcomes of the evaluation is largely lacking.

The focus of the S3 contest is on the evaluation of performance/scalability on the one hand and correctness/automation on the other. While the first is quantified with a number of runtime measures, the latter are compared by IR like measures. The recent cross-evaluation track of the S3 contest explicitly addresses coupling and to a certain degree (albeit rather informally) usability/effort. The functional scope considered is that of static discovery. The S3 contest does not take into consideration whether a framework could do more.

With respect to the evaluation criteria, the S3 contest fares similarly to the SWS Challenge concerning the utility and feasibility requirements. It has weaknesses regarding propriety and accuracy requirements. The latter is due mainly to the lack of reflection on the appropriateness and influence of the measures used. First steps to overcome this have been taken in the context of the cross-evaluation track [11].

1.5.3 The WS Challenge

The IEEE Web Service Challenge was founded in 2004 by M. Brian Blake of the University of Notre Dame. The challenge, itself, has been held annually since 2005. The event has been funded annually by the National Science Foundation.

The first event, initially named the IEEE EEE-05 Challenge, was organized by M. Brian Blake and William Cheung. While it started with evaluation of traditional web service frameworks and web service composition from a software engineering perspective, it has included composition via semantic services over the last several years. Evaluation measures used over time include the speed of the composition process, the correctness of the composition (measured in terms of accuracy, completeness, and minimal composition length), the execution time of the composed process (rewarding exploitation of parallelism), and the overall solution architecture.

An advantage of the WS Challenge compared, e.g., to the SWS Challenge is the unambiguous problem description provided. The WS Challenge has developed over two dozen different web service repositories from smaller, manually-created services (with realistic interfaces) to very large repositories with randomly generated semantic services. This approach is somewhat unique with respect to the other challenges. For the WS Challenge, it is also important to traverse a huge search space as efficiently as possible.

With respect to the evaluation dimensions, the WS Challenge measures a number of performance indicators and evaluates correctness of the algorithm, albeit not necessarily of the semantic reasoning. The functional scope is restricted to static composition. Coupling and usability/effort are not taken into account.

Regarding the criteria for good evaluations, the WS Challenge fares better than the other two when it comes to the accuracy requirements; it is almost comparable to them with regard to utility and propriety and is a bit less effective in determining the feasibility of real-life services. Since the initiative is a competition, sharing of approaches is less explicit than in the other challenges. Solutions are not generally developed through collaboration, but individual participants create their approaches separately and as a part of the forum techniques are discussed and perhaps incorporated individually for the next year. The other challenges seem to more encourage mutual understanding and learning.

The following parts will contain more detailed descriptions of the individual initiatives as well as experience reports from their participants. Most of the issues raised here will be touched upon in those chapters again.

1.6 The Future of the Initiatives

All three initiatives in which this book is based on are still running and are continuously being improved.

The S3 contest will continue to conduct annual events at least through 2013, with the existing tracks focused on the use of OWL-S and SAWSDL. It is anticipated that the OWL-S and SAWSDL test collections, which are used by the contest, will continue to grow and become further refined by the existing community effort that has been established.

The Semantic Web Service Challenge continues to be available online and has been collaborating with the SEALS project and running workshops in conjunction with the SEALS campaigns. As SEALS approaches the completion of its platform, there is also an opportunity to make the benchmarks from the SWS Challenge available in this platform. The SWS Challenge also counts on expanding its number of problem scenarios by contribution from the community.

The WS Challenge will continue to run annually. There are four newly anticipated aspects of the challenge in the coming years. The challenges will work on new dimensions of defining "what is good" with respect to quality of service. In previous challenges, all dimensions (performance, accuracy, efficiency, etc.) were treated equally. The challenge will apply weights in real time that competitors will need to acquire and leverage in their compositions. Also, the WS Challenge will incorporated dynamism in the service repositories. Instead of having a repository that is static throughout the evaluation process, we will remove and insert services in real time. This will prevent competitors from making one-time indexes for all challenge sets. The WS-Challenge will develop sets for security. Competitors will have to comply with a specific protocol in executing the sets. Finally, a challenge will be developed for service mashups. Instead of composing web services in workflows, the solution will be a set of services that creates mashups that are relevant to a specific purpose.

Regarding related initiatives that are contributing to a change in the evaluation landscape in the Semantic Web realm, the SEALS (Semantic Evaluation at Large Scale) project[11] is undertaking the task of creating a lasting reference infrastructure for semantic technology evaluation (the SEALS platform). The SEALS Platform will be an independent, open, scalable, extensible and sustainable infrastructure that will allow online evaluation of semantic technologies by providing an integrated set of evaluation services and a large collection of datasets. Semantic Web Services are one of the technologies which are supported by SEALS. The platform will support the creation and sharing of evaluation artifacts (e.g. datasets and measures) and services (e.g. retrieving data sets from repositories and automatic execution of tools), making them widely available according to problem scenarios, using semantic based terminology. A description of the results of SEALS for Semantic Web Service technologies and its relation with the current initiatives has been published in several deliverables (available from the website) and also in [4]. It is expected that the SEALS infrastructure, together with some of the outcomes of the project, such as a new dataset for WSMO-Lite [9], will benefit evaluation participants and organizers and advance the state-of-the-art of SWS evaluation.

[11] http://about.seals-project.eu/

1.7 Summary

We hope that we have convinced you by now – if you weren't from the beginning – that evaluations in computer science are important in order to advance the state of the art and to promote adoption of research results in real life applications. We have shown that this is also – or maybe even particularly – true for Semantic Web Services.

While necessarily short, we have introduced quality criteria that evaluations should meet and have compared existing evaluation initiatives for SWS with these criteria. We have also given a brief overview of the dimensions of evaluation for SWS and of which initiative addresses which of these dimensions.

Without having a closer look at the initiatives we can already conclude that while they do not meet all the criteria and do not address all dimensions equally, they offer a good starting point and are valuable.

In the next parts, you will find detailed reports on the initiatives supporting this view. There will be introductions by the organizers of the respective campaigns and in depth experience reports from participants.

We believe that one can learn three things from this book: It gives a good overview of existing approaches to SWS and discusses their respective weaknesses and strengths as found by the evaluations. It can thus serve as a guideline, if you are looking for a platform to use. Second, it gives a detailed overview of the state of the art in evaluation of SWS. If you are a developer of an SWS framework, the dimensions discussed and the experiences made by participants in the evaluation campaigns might guide you towards improvements of your solution. Better yet, of course, take part in one of the campaigns yourself. This book will help you identify the one that addresses the issues that you are most concerned about. Third, we hope to contribute to the progress of evaluation in computer science in general. This book should give a good impression on what to consider when planning an evaluation campaign and which results to expect.

References

1. T. Berners-Lee, J. Hendler, O. Lassila, The semantic web. Sci. Am. **284**(5), 34–43 (2001)
2. W. Beywl, Selected comments to the standards for evaluation of the german evaluation society – English edition. Technical report, German Evaluation Society (DeGEval), (2003)
3. M.B. Blake, H. Gomaa, Agent-oriented compositional approaches to services-based cross-organizational workflow. Decis. Support Syst. **40**(1), 31–50 (2005)
4. L. Cabral, I. Toma, Evaluating semantic web services tools using the SEALS platform, in *Proceedings of IWEST Workshop at ISWC 2010*, Shanghai, 2010
5. J. Domingue, L. Cabral, S. Galizia, V. Tanasescu, A. Gugliotta, B. Norton, C. Pedrinaci, IRS-III: a broker-based approach to semantic web services. J. Web Semant. **6**(2), 109–132 (2008)
6. N. Fenton, S.L. Pfleeger, R.L. Glass, Science and substance: a challenge to software engineers. IEEE Softw. **11**(4), 86–95 (1994)

7. G. Gediga, K.-C. Hamborg, I. Düntsch, Evaluation of software systems, in *Encyclopedia of Computer Science and Technology*, vol. 45, ed. by A. Kent, J.G. Williams (Marcel Dekker, Inc., CRC, New York, 2002), pp. 127–153
8. D. Harman, Overview of the first Text REtrieval Conference (TREC-1), in *Proceedings of the First Text REtrieval Conference (TREC-1)*, Gaithersbury, 1992
9. J. Kopecky, T. Vitvar, WSMO-Lite: lowering the semantic web services barrier with modular and light-weight annotations, in *Proceedings of 2nd IEEE International Conference on Semantic Computing (ICSC)*, Santa Clara, CA, USA, 2008
10. U. Küster, An evaluation methodology and framework for semantic web services technology, University of Jena, Berlin Logos, 2010
11. U. Küster, B. König-Ries, Relevance judgments for web services retrieval-a methodology and test collection for sws discovery evaluation, in *2009 Seventh IEEE European Conference on Web Services* (IEEE, Los Alamitos, 2009), pp. 17–26
12. U. Küster, B. König-Ries, C. Petrie, M. Klusch, On the evaluation of semantic web service frameworks. Int. J. Semant. Web Inf. Syst. **4**(4), 31–55, (2008)
13. H. Lausen, C. Petrie, M. Zaremba, W3C SWS testbed incubator group charter (2007), Available online at http://www.w3.org/2005/Incubator/swsc/charter
14. D. Martin, M. Burstein, D. McDermott, S. McIlraith, M. Paolucci, K. Sycara, D.L. McGuinness, E. Sirin, N. Srinivasan, Bringing semantics to web services with OWL-S. World Wide Web **10**(3), 243–277 (2007)
15. C. Petrie, U. Küster, T. Margaria-Steffen, W3C SWS challenge testbed incubator methodology report. W3c incubator report, W3C, (2008), Available online at http://www.w3.org/2005/Incubator/swsc/XGR-SWSC/
16. W.F. Tichy, Should computer scientists experiment more? IEEE Comput. **31**(5), 32–40 (1998)
17. M.V. Zelkowitz, D.R. Wallace, Experimental models for validating technology. Computer **31**(5), 23–31 (1998)

Part I
Results from the S3 Contest: OWL-S and SAWSDL Matchmaker Evaluation Tracks

Chapter 2
Overview of the S3 Contest: Performance Evaluation of Semantic Service Matchmakers

Matthias Klusch

Abstract This chapter provides an overview of the organization and latest results of the international contest series on semantic service selection (S3). In particular, we introduce its publicly available S3 evaluation framework including the standard OWL-S and SAWSDL service retrieval test collections OWLS-TC and SAWSDL-TC as well as its retrieval performance evaluation tool SME2. Further, we classify and present representative examples of Semantic Web service matchmakers which participated in the S3 contest from 2007 to 2010. Eventually, we present and discuss selected results of the comparative experimental performance evaluation of all matchmakers that have been contested in the past editions of the S3 series.

2.1 Introduction

In the rapidly growing Internet of services, efficient means for service discovery, that is the process of locating existing services based on the description of their (non-)functional semantics are essential for many applications. Such discovery scenarios typically occur when one is trying to reuse an existing piece of functionality (represented as a Web service) in building new or enhanced business processes. Matchmakers [6] are tools that help to connect a service requestor with the ultimate service providers. The process of service selection or matchmaking encompasses (a) the pairwise semantic matching of a given service request with each service that is registered with the matchmaker, and (b) the semantic relevance ranking of these services. In contrast to a service broker, a service matchmaker only returns a ranked list of relevant services to the requestor together with sufficient provenance

M. Klusch (✉)
German Research Center for Artificial Intelligence (DFKI), Stuhlsatzenhausweg 3, Saarbruecken, Germany
e-mail: klusch@dfki.de

M.B. Blake et al. (eds.), *Semantic Web Services*, DOI 10.1007/978-3-642-28735-0_2,

information that allows to directly contact the respective providers. A matchmaker neither composes nor negotiates nor handles the execution of services.

Semantic matching of services determines the degree of semantic correspondence between the description of a desired service, that is the service request, and the description of a registered service, that is the service offer. For this purpose, both service request and service offer are assumed to be described in the same format. In this chapter, we focus on semantic service matchmakers [5] that are capable of selecting semantic services in formats such as OWL-S,[1] SAWSDL[2] or WSML,[3] that is services whose functionality is described by use of logic-based semantic annotation concepts which are defined in one or multiple formal ontologies [4]. The processing of such semantic annotations for service selection by a matchmaker bases either on a global ontology it is assumed to share with service consumers and providers, or on the communication of sufficient ontological information on service annotation concepts to the matchmaker for this purpose. The performance of any service matchmaker can be measured in the same way as information retrieval (IR) systems are evaluated for decades, that is in terms of performance measures like recall, average precision and response time.

Though many implemented semantic service matchmakers exist, there was no joint initiative and framework for the comparative experimental evaluation of their retrieval performance available until a few years ago. For this reason, the international contest series on semantic service selection (S3) has been initiated in 2006 by DFKI together with representatives of several other institutions and universities in Europe and USA. Since then it has been organized annually based on a publicly available S3 evaluation framework for semantic service selection which actually consists of the standard test collections OWLS-TC[4] and SAWSDL-TC,[5] as well as the evaluation tool SME2.[6] The participation in the contest is by online submission of a matchmaker plugin for the SME2 tool while the final results of each edition of the contest are presented at a distinguished event such as at a major conference of the Semantic Web or relevant community and/or on the official Web site of the S3 contest.[7] The S3 contest series has been exclusively funded by the German ministry of education and research (BMB+F) under project grants 01IW08001 (MODEST, http://www.dfki.de/-klusch/modest/) and 01IW08005 (ISReal, http://www.dfki.de/-klusch/isreal/).

The remainder of this chapter is structured as follows. We briefly introduce the S3 evaluation framework in Sect. 2.2. This is followed by a classification of all participants of the contest from 2007 to 2010 together with brief descriptions of

[1] http://www.w3.org/Submission/OWL-S/

[2] http://www.w3.org/2002/ws/sawsdl/

[3] http://www.wsmo.org/wsml/wsml-syntax

[4] http://projects.semwebcentral.org/projects/owls-tc/

[5] http://projects.semwebcentral.org/projects/sawsdl-tc/

[6] http://projects.semwebcentral.org/projects/sme2/

[7] http://www.dfki.de/-klusch/s3/