Evaluation of Semantic Service Discovery - A Survey and Directions for Future Research

Ulrich Küster[1], Holger Lausen[2], and Birgitta König-Ries[1]

[1] Institute of Computer Science, Friedrich-Schiller-University Jena, 07743 Jena, Germany, ukuester|koenig@informatik.uni-jena.de
[2] Digital Enterprise Research Institute, University of Innsbruck, 6020 Innsbruck, Austria, holger.lausen@deri.at

Abstract. In recent years a huge amount of effort and money has been invested in the area of semantic service discovery and presented approaches have become more sophisticated and mature. Nevertheless surprisingly little effort is being put into the evaluation of these approaches. We argue that the lack of established and theoretically well-founded methodologies and test beds for comparative evaluation of semantic service discovery is a major blocker of the advancement of the field. To lay the ground for a comprehensive treatment of this problem we discuss the applicability of well-known evaluation methodologies from information retrieval and provide an exhaustive survey of the current evaluation approaches.

1 Introduction

In recent years semantic services research has emerged as an application of the ideas of the semantic web to the service oriented computing paradigm. Semantic web services (SWS in the following) have received a significant amount of attention and research spending since their beginnings roughly six years ago [1]. Within the sixth EU framework program[1] (which ran from 2002 to 2006) alone at least 20 projects with a combined funding of more than 70 million Euro deal directly with semantic services which gives a good impression of the importance being currently put on this field of research. In the following we will focus on efforts in the field of SWS discovery and matchmaking. We refer to discovery as the whole process of retrieving services that are able to fulfill a need of a client and to matchmaking as the problem to automatically match semantically annotated service offers with a semantically described service request. However, we think that our findings also apply to other related areas, like automated semantic service composition.

In this paper we argue that despite of the huge amount of effort (and money) spent into SWS discovery research and despite the fact that the presented approaches become more sophisticated and mature, much too little effort is put into the evaluation of the various approaches. Even though a variety of different

[1] http://cordis.europa.eu/fp6/projects.htm

service matchmakers have been proposed we did not suceed to find any publications with a thorough, systematic, objective and well designed evaluation of those matchmakers. This corresponds to a trend that seems to exist in computer science in general. In [2] Tichy et. al. find that computer scientists publish relatively few papers with experimentally validated results compared to other sciences. In a follow-up work [3] Tichy claims that this trend is harmful for the progress of the science.

There are positive examples that back his claim:

> "[in the experiments] ... there have been two missing elements. 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 importance of this is not to show any system to be superior, but to allow comparison across a very wide variety of techniques, much wider than only one research group would tackle. [...] The second missing element, which has become critical [...] is the lack of a realistically-sized test collection. Evaluation using the small collections currently available may not reflect performance of systems in large [...] and certainly does not demonstrate any proven abilities of these systems to operate in real-world [...] environments. This is a major barrier to the transfer of these laboratory systems into the commercial world."

This quote by Donna Harman [4] addressed the situation in text retrieval research prior to the establishment of the series of TREC conferences[2] in 1992 but seems to perfectly describe the current situation in SWS discovery research. Harman continued:

> "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."

From the perspective of today, it is clear that her hope regarding the positive influence of the availability of mature evaluation methods to the progress of information retrieval research was well justified. In this paper we argue that a similar effort for SWS related research is necessary today for the advancement of this field.

The rest of this paper is organized as follows. In Section 2 we will review the philosophy of information retrieval evaluation and argue that traditional evaluation methods can not be applied easily to the domain of SWS matchmaking. In Section 3 we provide an extensive survey of current evaluation efforts in the area of SWS discovery. We cover the related work in Section 4, draw conclusions about what is missing so far and provide directions for future work in Section 5 and finally summarize in Section 6.

2 The Philosophy of Information Retrieval Evaluation

In a broader context, discovery of semantic web services can be seen as a special information retrieval (IR) problem. According to Voorhees [5], IR evaluation has

[2] http://trec.nist.gov/

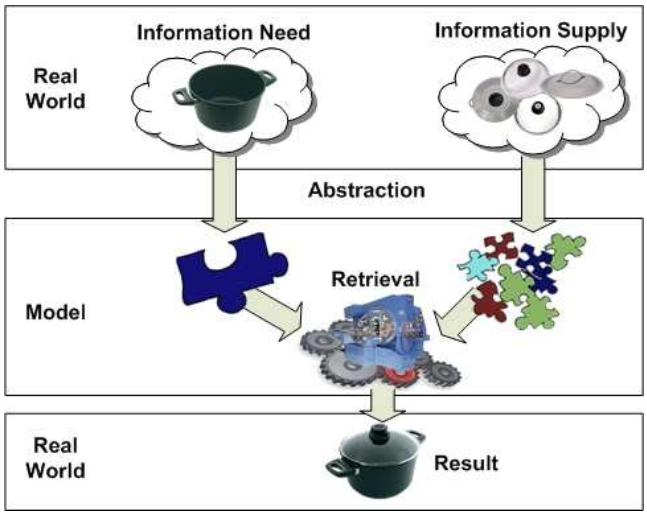

Fig. 1. The process of information retrieval

been dominated for four decades by the Cranfield paradigm which is characterized by the following properties:

- An IR system is mainly evaluated by means of *recall* and *precision*.
- Recall is defined as the proportion of retrieved documents that are *relevant*.
- Precision is defined as the proportion of *relevant* documents that are retrieved.
- Relevance is based on topical similarity as obtained from the judgements of domain experts.
- Test collections therefore have three components: a set of documents (the test data), a set of information needs (topics or queries) and a set of relevance judgements (list of documents which should be retrieved)

Vorhees identifies several assumptions on which the Cranfield paradigm is based that are unrealistic in most cases. She concludes that experiments based on those assumptions are a noisy process but is able to provide evidence that – despite of the noise – such experiments yield useful results, as long as they are only used to assess the *relative performance* of different systems evaluated by the *same experiment*.

Since the experiments based on the Cranfield paradigm are extremely well established, since the methodology of these experiments is well understood and since SWS matchmaking is a special information retrieval problem, it seems obvious to try to apply the same methods and measures to the SWS matchmaking domain. However, in the following we argue that this is not a promising approach for various reasons.

A model of the general process of information retrieval is depicted in Figure 1. The user has a real world need (like information about a certain topic) that needs

to be satisfied with the existing real world supply (like a collection of documents). Both the need and the supply are abstracted to a model. In the case of web search engines for instance, such a model will consist of descriptors extracted from the query string and data structures like indexes built upon descriptors extracted from the web pages etc. The information retrieval system then operates on that model to match the need with the supply and returns the (real world) results. As a matter of fact the power of this model (how well it captures the real world and how well it supports the retrieval, i.e. matchmaking and ranking process) is of critical importance for the retrieval system and thus a central component of its overall performance.

Traditional information retrieval systems typically create the model they operate on in an autonomous fashion. Thus, from the viewpoint of an evaluation they operate on the original data. Consequently, completely different IR systems can be evaluated on a common test data set (like a collection of documents).

SWS matchmaking follows a different paradigm. Here the semantic annotation is the model that is exploited during the matchmaking and it is not created automatically, but written by human experts. Currently there is no agreed upon formalism used for the semantic annotations, but competing and mostly incompatible formalisms are in use (like WSMO[3], OWL-S[4], WSDL-S[5], DSD[6], ...).

To apply the Cranfield paradigm to the evaluation of SWS discovery, one could provide a test collection of services in a particular formalism (e.g. OWL-S) and limit participation in the experiment to systems based on that formalism. This is the approach of the current S3 Matchmaker Contest (Section 3.1). Unfortunately, this excludes the majority of systems from participation. But there is an even more severe issue. Virtually all semantic matchmaking systems that are based on some form of logical reasoning operate deterministically. Therefore the question whether a *semantic* offer description matches a *semantic* request description in the same formalism can usually be decided unambiguously yielding perfect recall and precision (only depending on the definition of *match*). The task to evaluate, however, is the retrieval of *real-world* services that match a *real-world* request. Thus, the major source of differentiation between various approaches is the expressivity of the employed formalism and reasoning. The critical questions here are:

- How precisely can a description based on a particular formalism reflect the real-world semantics of a given service (offer or request)?
- How much of the information contained in the descriptions of a service can a matchmaker use efficiently and effectively to decide whether two services match?

Note that the first question usually calls for more expressive formalisms whereas the second one requires less expressive formalisms to keep the reasoning tractable.

[3] http://www.wsmo.org/
[4] http://www.daml.org/services/owl-s/
[5] http://lsdis.cs.uga.edu/projects/meteor-s/wsdl-s/
[6] http://hnsp.inf-bb.uni-jena.de/DIANE/

As argued above, the formalism employed for the semantic annotation of the services, i.e. the model used for the matchmaking, is of crucial importance for the overall performance of the discovery system. Consequently, a good evaluation of SWS discovery should measure not only the performance of a system for a given formalism, but also evaluate the pros and cons of that formalism itself. In fact, as long as there is no common understanding about the pros and cons of different formalisms and no agreement about which formalism to employ for a given task, evaluation of SWS discovery should first and foremost help to establish this missing common understanding and agreement.

The approach outlined above, however, neglects the influence of the model used in the matchmaking process and therefore does not measure the performance of that part of the retrieval process, which has the largest influence to the overall performance of the system.

A different approach that overcomes this limitation would be to provide a test collection of services in any format (e.g. human language) and let the participants annotate the services manually with their particular formalism. This is the approach taken by the SWS-Challenge (Section 3.2) and the DIANE evaluation (Section 3.3)). Unfortunately there are problems with this proceeding, too. First, such an experiment can hardly be performed on a large test collection, since the effort for the participants to manually translate the services into their particular formalisms is enormous. Yet, the unavoidable noise of experiments based on the Cranfield paradigm precisely requires large test collections to yield stable results [5]. Second, due to the human involvement, such an experiment can not be conducted in an automated way. Even worse, such an experiment does not only measure the performance of the matchmaking formalism and system, but also the abilities of the experts that create the semantic annotations. This introduces a whole new dimension of noise to the evaluation.

For the reasons given above we conclude that the experimental setup and the evaluation measures and methods developed for traditional information retrieval do not transfer directly to the SWS discovery domain. The influence of the described problems needs to be explored and new methods and measures have to be developed where necessary. To lay the foundation for this task, we provide an extensive survey of the current efforts in SWS discovery evaluation in the following section.

3 A Survey Of Current Approaches in Semantic Web Service Discovery Evaluation

3.1 S3 Matchmaker Contest

Klusch et al. have recently announced an annual international contest S3 on Semantic Service Selection[7] whose first edition will be held in conjunction with the upcoming International Semantic Web Conference in Busan, Korea (November 2007). We would like to express our acknowledgement and appreciation of

[7] http://www-ags.dfki.uni-sb.de/~klusch/s3/

this new effort that we welcome very much. Despite of that we identify some problems in the current setup of the contest.

The contest is based on a test collection of OWL-S services and "evaluation of semantic web service matchmakers will base on classic performance metrics recall/precision, F1, average query response time"[7]. As argued in the previous section this style of application of the Cranfield paradigm to the SWS matchmaking domain has a limited scope and significance since it does not allow a comparative evaluation of different semantic formalisms.

We furthermore think there is currently a problematic flaw in the practical setup of the contest, too. The most severe achilles' heel of any such contest is the dependency on a good SWS test collection. This year the S3 contest will rely solely upon the OWL-S Test Collection 2[8] which we believe to be unsuitable for a meaningful comparative and objective SWS matchmaking evaluation. We will explain our skepticism by a critical review of the collection.

The OWL-S Test Collection 2[9] is the only publicly available test collection of semantically annotated services of mentionable size. It has been developed within the SCALLOPS project[10] at the German Research Centre for Artificial Intelligence (DFKI). The most recent version 2.1 of the collection (OWLS-TC2 released in October 2006) contains 582 semantic web services written in OWLS 1.1. To put our following criticism into the correct light and in acknowledgement that currently no better public standard test collection exists, we would like to mention that the OWLS-TC2 does not claim to be more than "one possible starting point for any activity towards achieving such a standard collection by the community as a whole"[6]. Our criticism of OWLS-TC2 covers three aspects.

Use of realistic real-world examples. One common criticism to many use cases and evaluations in the service matchmaking domain is the use of artificial toy examples which are far from realistic applications. Even though examples do not necessarily have to be realistic to test features of a matchmaking system, the use of real-world examples clearly minimizes the danger of failing to detect lacking features or awkward modeling. Furthermore, toy examples far from real-world applications critically hinder the acceptance of new technology by industry. OWLS-TC2 claims that "the majority of [...] services were retrieved from public IBM UDDI registries, and semi-automatically transformed from WSDL to OWL-S"[6]. Thus, one would expect somewhat realistic services but a substantial share of the 582 services of OWLS-TC2 seems quite artificial and idiosyncratic. Oftentimes the semantic of the service is incomprehensible even for a human expert and unfortunately only six of the original WSDL files are included in the test set download. A comprehensive coverage is impossible due to the size of OWLS-TC2 but the following examples illustrate the issues (in the following

[8] It is planned to extend the scope of the contest beyond OWL-S based matchmakers in the future. However, public test collections based on other formalisms have unfortunately not been developed so far. The S3 contest organizers have set up a public wiki (http://www-ags.dfki.uni-sb.de/swstc-wiki) to initiate efforts in this direction.

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

[10] http://www-ags.dfki.uni-sb.de/~klusch/scallops/

service names always refer to the name of the corresponding service description file, not the service name from the service's profile, quotes are from the service's description file or the OWLS-TC2 manual):

- Some services are simply erranous, quite a few services for instance are pairwise identical except for the informal textual description (e.g. `_price_CannonCameraservice.owls` and `_price_Fishservice.owls`)
- The service `__destination_MyOfficeservice.owls` is supposed to "return destination of my office", but takes concepts of type *organization* and *surfing* (which is a subclass of *sports*) as input.
- The service `surfing_farmland_service.owls` is described as "This is the recommended service to know about the farmland for surfing" and has an input of type *surfing* and an output of type *farmland*. What's the semantic of this service?
- The service `qualitymaxprice_cola_service.owls` "provides a cola for the maximum price and quality. The quality is an optional input." It is described by its inputs of type *maxprice* and *quality* and an output of type *cola*. There are a whole lot of similar services that return cola (six more services), beer + cola, coffee + whiskey (eleven services), cola-beer, cola + bread or biscuit (two services), drinks (three services), liquid, whiskey + cola-beer as well as irish coffee + cola. It remains unclear what is the semantics of these services.

Besides these issues we believe that examples from domains like *funding of ballistic missiles*, which the typical user of an evaluation system does not have any experience with, make a realistic evaluation unnecessary difficult.

Semantic richness of descriptions. Services should not only be realistic and realistically complex, they also should be described in sufficient detail to allow for meaningful semantic discovery. After all there should be an advantage to use semantic annotations compared to simply using traditional information retrieval techniques. Unfortunately the services of OWLS-TC2 are described extremely superficial. First of all it seems that all services are solely described by their inputs and outputs. What is the semantic of a service (`car_price_service.owls`) that takes a concept of type *Car* as input and has a concept of type *Price* as output? It might sell you a car and tell you the price afterwards, it might just as well only inform you about the price of a new car or the price of a used car. It might rent a car for the returned price. It might tell you the price of the yearly inspection for the given car. There are many different possible interpretations. What is the semantic of a service like `car_priceauto_service.owls` that takes as input a concept of type *Car* and has outputs of type *Price* and *Auto* (which is a subclass of car)?

In our view the services of OWLS-TC2 are not described in sufficient detail to allow to perform meaningful semantic discovery on them. The problem is greatly aggravated by the fact that the services in OWLS-TC2 make use of classes in a class hierarchy but do not make use of attributes or relations. Thus, in most cases the semantic of the services is greatly underspecified and - if at all -

understandable only from the informal textual documentation[11]. Overall it seems the textual descriptions of the service offers and queries are not captured well by the semantic descriptions. Query 23 for instance is informally described as "the client wants to travel from Frankfurt to Berlin, that's why it puts a request to find a map to locate a route from Frankfurt to Berlin." This request is described (`geographicalregiongeographical-region_map_service.owls`) as a request for a service with two unordered inputs of type *geographical region* and a single output of type *map*. Clearly routing services will also be found (among many others) by such a request, but we are afraid that offers and requests described at this level of detail will neither allow to demonstrate the added value of *semantic* service discovery nor to evaluate the power of matchmakers which should create this added value.

Independence of offer and request descriptions Ideally, service offer and request descriptions should be designed independently since this is the envisioned situation in reality. Service providers describe their offers, clients query for a service with a semantic request description and a matchmaker is supposed to find the offers that match the request. We acknowledge that in laboratory settings it is sometimes desirable to artificially design the offers to match a request at various degrees. This way it can be assured that all potentially existing degrees of match occur during a test run. However, a test where the offers have been designed to match a request at hand with specific degrees runs the risk of doing nothing more than supporting the belief that a particular matchmaker *implementation* operates as expected. It does not demonstrate the power of a certain semantic description formalism or a certain matchmaking approach. Despite the fact that OWLS-TC2 claims that most services where retrieved from public IBM UDDI registries, we got the impression that for most of the queries in OWLS-TC2 the matching services have been artificially designed for that particular query. Query 4 for instance asks for the combined price of a car and a bicycle. It seems quite idiosyncratic to buy a car and a bicycle as a package, yet there are at least eleven service offers in OWLS-TC2 that precisely offer to provide the price of a package of one car and one bicycle. Our impression is further backed up by the fact that the number of relevant services is quite stable for all the queries.

Conclusions OWLS-TC2 has been developed by the effort of a single group to evaluate a particular (hybrid) matchmaker [7] and the OWLS-TC2 manual states that it has been designed to be balanced with respect to the matching filters of that matchmaker, i.e. besides performing semantic discovery it explicitly also uses classical Information Retrieval techniques. Thus OWLS-TC2 is suited to test and evaluate the features of this particular hybrid matchmaker, but for the reasons given above we do not think this test collection is suited for a broader comparative evaluation of different semantic matchmakers. Based on this finding

[11] This may have been on purpose since the OWL-MX matchmaker, the matchmaker OWLS-TC was designed for, is a hybrid matchmaker that combines semantic matchmaking with traditional information retrieval techniques

and the discussion in Section 2 we doubt that the current setup of the S3 contest will yield meaningful results.

To put our criticism above into the correct context, we would like to acknowledge once more that sadly there is currently no better public test collection than OWLS-TC2 and that the creation of a balanced, realistic and rich, high-quality semantic service test collection involves an immense amount of effort that clearly exceeds the capabilities of any single group. The organizers of the S3 Contest have therefore stressed that such a collection can only be built by the community as a whole and that the contest and its current employment of OWLS-TC2 is only a first step in that direction. They have set up a wiki[12] to initiate a corresponding community effort. We hope that our critical analysis of OWLS-TC2 will help to motivate such community effort and will therefore ultimately help to improve the quality of the emerging collections.

3.2 Semantic Web Service Challenge

The Semantic Web Service Challenge is an initiative aiming to create a test bed for frameworks that facilitate the automation of web service mediation and discovery. It is organized as a series of workshops in which participants try to model and solve problems described in the publicly available test bed. The test bed is organized in scenarios (e.g. discovery or mediation), each one containing detailed problem descriptions. Compared to the S3 contest the number of available services (at the time of writing around a dozen) is relatively small, however the organizers put strong emphasis on providing realistic and detailed scenarios.

The Challenge organizers have realized that the lack of comprehensive evaluation and test beds for semantic web service system is one of the major blockers for industrial adoption of the used techniques. They have designed the challenge having the following ideas in mind:

- *Solution Independence.* Existing test cases often suffer from the problem that they have been reverse engineered from the solution, i.e. that the use case has been created according to the strengths of a particular solution. This hinders comparison across multiple systems. By letting the organizers not directly participate and by defining rules on how new scenarios can be added the SWS Challenge tries to overcome this problem.
- *Language Neutral.* Closely connected to the above issue is the one how to describe the problem set. Using a particular formalism for describing services already implies the solution to a huge degree. In our opinion, the choice of the right level of detail to include in the service and goal descriptions in fact still constitutes one of the core research problems and should not be dictated by the test bed for an evaluation. The SWS Challenge organizers have consequently decided not to provide formal descriptions but only natural language ones.

[12] http://www-ags.dfki.uni-sb.de/swstc-wiki

– *No Participation Without Invocation.* Each scenario provided comes with a set of publicly available web services. On the one hand this should yield in some industrial relevance, on the other hand it provides the organizers with an unambiguous evaluation method. If a system claims to be able to solve a particular problem (e.g. discovery of the right shipment provider), this can be automatically verified by monitoring the SOAP messages exchanged.

Scenario Design. Within the research community only little consensus exists about what information should be included in a static service description and how they should be semantically encoded. The scenarios are thus described using existing technologies (WSDL, XSD, and natural language text descriptions). In the following we will explain the philisophy of the scenarios by means of the first discovery scenario[13] provided. This scenario includes five shipment services that are modeled according to the role models of order forms of existing shipment companies on the Internet. The services are backed by corresponding implementations that are part of the test bed.

The task to solve is to discover and invoke a suitable shipper for a given shipping request. The scenario contains a set of such requests which are categorized into levels of increasing difficulty. It starts with *Discovery Based on Destination* and adds weight and price criteria as well as simple composition and temporal constraints to the more difficult problems. For each request the problem description contains the expected correct solution (i.e. the list of matching services) already.

Evaluation Methodology. Solutions to a scenario are presented at the SWS-Challenge workshops. The evaluation is performed by teams composed of the workshop organizers and the peer participants. The organizers are aware, however, that this causes scalability problems if the number of participants increases and also is not strictly objective.

The evaluation approach focuses on evaluating the functional coverage, i.e. on whether a particular level of the problem could be solved by a particular approach or formalism correctly or not. The intention is to focus on the *how*, that is the concrete techniques and descriptions an approach uses to solve a problem and not on the time it requires for execution, thus no runtime performance measurements are taken.

The organizers argue that in practice automatic and dynamic discovery is not widely used, thus part of the challenge is to refine the challenge and to illustrate the benefit of using semantic descriptions. The basic assumption to test is whether approaches which rely more heavily on semantic annotations will be easier adaptable to changes in the problem scenarios. Therefore the challenge does not only certify functional coverage, but initially it was planned to also assess on how *elegant* a solution can address the problems posed and how much effort was needed to proceed from a simpler to a more complex problem level.

However it turned out that it is extremely difficult to assess this in an objective manner [8]. Measurements based on counting the number of lines (or

[13] http://sws-challenge.org/wiki/index.php/Scenario:_Shipment_Discovery

statements) of semantic description do not adequately represent the usability of an approach. Also the measurements of changes that are required to solve new problems turned out to be problematic. They worked in the beginning when all participants started with the same known set of problem levels which was then extended at consecutive workshops. However, participants entering the challenge right now have access to all problem levels right away which makes an objective assessment of the necessary change to solve the more advanced levels on top of the simpler ones impossible. It is planed to test the usability of surprise scenarios for the envisioned assessment at the next workshop.

Lessons Learned. By only describing the problems without any particular formalism in mind the SWS Challenge organizers where able to attract various different teams from different communities. Thus it successfully enables evaluation across very heterogenous solutions. By requiring a grounding in real web services a significant amount of effort was consumed both on the site of the organizers as well of the participants with problems related to standard web service technology, which are not strictly relevant when looking at discovery in isolation. This also may have discouraged potential teams more familiar with knowledge representation than with web service technology. On the other site the implementation has been proven useful to (1) disambiguate the natural language text descriptions and (2) undoubtedly show whether a participant has or has not solved a particular problem. By having an implementation, no one could change the scenario to fit their solution without failing at the automated tests based on exchanged SOAP messages.

With respect to the scenarios being described in informal natural language only, it turned out that the original scenarios were indeed ambiguous in several cases. However, during the course of usage of a particular scenarios these ambiguities where discovered by the participants and could subsequently be resolved by the authors of the particular scenario. Our experience shows that this way even scenarios described in natural language only become sufficiently well-defined over time. Usually the implementation also does disambiguate a scenario, however it is not the most efficient way to find out about the intention of a particular aspect.

3.3 DIANE Service Description Evaluation

Within the DIANE project[14], a service description language, DIANE Service Description (DSD) and an accompanying middleware supporting service discovery, composition, and invocation have been developed. DIANE is one of the projects taking part in the SWS Challenge. Besides the evaluation provided by the challenge, considerable effort has been put into devising an evaluation suite for semantic service description languages[9]. While this work is certainly not completed yet, it complements the SWS Challenge in some important aspects. The DIANE evaluation focuses on four criteria an evaluation should measure:

[14] http://hnsp.inf-bb.uni-jena.de/DIANE/

1. *Degree of Automation:* Are the language and the tools powerful enough to allow for automatic and correct service usage? That means: Given a service request and service offers, will the discovery mechanism find the best-matching service offer and will it be possible to automatically invoke the service based on these results?
2. *Efficiency of Matchmaking:* Is it possible to efficiently and correctly compute the matchvalue of arbitrary offers and requests?
3. *Expressiveness:* Is it possible to describe real services and real service requests in sufficient detail to meet Criteria 1? Can this be done with reasonable effort?
4. *Decoupling:* Will a discovery mechanism be able to determine similarity between service offers and requests that are developed independently of each other? In other words: If a service requester writes his request without knowledge of the existing service descriptions, does the language offer enough guidance to ensure that suitable and only suitable services will be found by the discovery mechanism?

It is quite obvious, that these criteria require contradictory properties of the description language: While, e.g., Criterion 3 requires a highly expressive language, Criterion 2 will be the easier to meet the less powerful the language is. Service description languages thus need to strike a balance between these competing requirements.

Criteria 1 and 2 can be evaluated basically by providing a proof-of-concept implementation. We will not look at them in more detail here but instead focus on the more interesting Criteria 3 and 4. To evaluate these, a benchmark has been designed. This benchmark has been used for the DSD evaluation, so far. It is, however, not language specific and can be used for other approaches, too.

To evaluate how well a service description language meets Criterion 3, a set of real world services is needed. As mentioned earlier in the paper, the example of the OWL-S TC shows that meaningful real world services are apparently not easy to come by. In particular, meaningful real world services that are described in sufficient detail are scarcely available. For our benchmark, we therefore chose a different approach: A group of test subjects not familiar with semantic web technology were asked to formulate service requests for two different application domains. We have chosen a bookbuying and train ticket scenario with typical end user requests as one domain and a travel agency looking for external services that can be included in applications as the second domain. The queries the test subjects devised were formulated in natural language. This resulted in about 200 requests. In preparation of the benchmark, domain experts developed ontologies they deemed necessary to handle the two domains (books, money, trains, ...). Subsequently, the experts attempted to translate the requests into DSD and computed how many requests could be directly translated, how many could be translated but required extensions of the ontologies and how many could not be appropriately expressed using the language constructs provided by DSD. These three values measure how well the language is able to describe realistic services of different types.

To evaluate whether decoupled description of offers and requests is possible, a number of the test subjects were given an introduction to DSD. They were subsequently divided into two groups that were not allowed to communicate with each other. The groups were then asked to formulate service offers and requests, respectively, from a given natural language description. The resulting DSD description were then evaluated by the matcher and precision and recall of the matchmaking were determined. High values for both parameters indicate that it is indeed possible to decouple offer and request description. A summary of the benchmark queries and results can be found online[15].

3.4 Other Approaches

The annual IEEE Web Service Challenge[16] [10] is similar in spirit to the S3 Matchmaker Contest, but focussed rather on syntactic or low level semantic matchmaking and composition based on matching WSDL part names whereas we focus on explicit higher level semantics.

Toma et al. [11] presented a framework for the evaluation of semantic matchmaking frameworks by identifying different aspects of such frameworks that should be evaluated: query and advertising language, scalability, reasoning support, matchmaking versus brokering and mediation support. They evaluate a number of frameworks in the service as well as the grid community with regard to these criteria. The focus of the work is rather on the excellent survey than on the comparison framework itself. While the framework does provide guidance for a structured comparison, it does not offer concrete test suites, measures, benchmarks or procedures for an objective comparative evaluation.

In her PhD thesis [12], Åberg proposes a platform to evaluate service discovery in the semantic web. However, her platform is rather a software architecture to provide some guidance in the development of SWS frameworks than a real evaluation platform and it does not become clear how this platform can help to comparatively evaluate different web service frameworks. Despite of some interesting starting points she does not provide a comprehensive framework (neither in theory nor practice) that can be used for the evaluation of different discovery approaches and furthermore ignores related approaches (like the SWS-Challenge or the S3 Matchmaker Contest) completely.

Moreover we have looked into the evaluation results of various SWS research projects (see for instance [13–15]). Many have spent a suprisingly small share of resources on evaluation. For example RW^2, an Austrian funded research project[17], has implemented different discovery engines for request and service description in different logical languages, respectively different granularity. However as evaluation only a relatively small set of a couple of dozen handcrafted services exist. The EU projects DIP and ASG have also developed similar discovery engines. With respect to evaluation they quote industrial case studies,

[15] http://hnsp.inf-bb.uni-jena.de/DIANE/benchmark/
[16] http://www.ws-challenge.org/
[17] http://rw2.deri.at/

	S3 Contest	SWS-Challenge	DIANE
Scope of evaluation			
Runtime performance	+	–	–
Framework tool support and usability	–	–	o
Expressivity of formalism and matchmaking	–	+	+
Supported level of decoupling	–	–	+
Quality of evaluation			
Neutral to formalism	o	+	+
Independent from solution	–	+	o
Realistic and complex use cases	–	+	o
Large test set	+	–	o

Table 1. Preliminary comparison of complementary strengths of the existing efforts.

however, in essence those are also just a small set of service descriptions. Moreover due to intellectual property rights restrictions the situation is even slightly worse, since not all descriptions are publicly available and a comparative evaluation is thus impossible.

3.5 Conclusions

Our survey has shown that - despite of the amount of attention that SWS discovery and matchmaking research receives - surprisingly little effort is devoted to experimental and comparative evaluation of the various approaches. We found only three approaches that intensively deal with SWS discovery evaluation in particular. Table 1 shows a schematic and simplified comparison of the different strengths of these approaches which is only meant to give a very high level summary of the extensive treatment above. All approaches can only be seen as starting initiatives in the right direction. The SWS-Challenge is currently the best established initiative in the field, but the S3 Matchmaker Contest and the DIANE evaluation complement it in important aspects. In particular the notions of *decoupling* the creation of offer and request descriptions, of involving *inexperienced users* in devising descriptions and all aspects related to *runtime performance* comparisons are not covered by the challenge so far.

4 Related Work

The existing approaches to evaluate SWS discovery have been covered extensively above. However, these approaches have not provided a critical review of the evaluation process itself so far. In contrast, the evaluation of evaluation in traditional information retrieval has been subject of a number of studies, e.g. by Saracevic [16] or Voorhees [5], but as we argue in Section 2, the results can not be directly applied to the domain of SWS discovery. Similar meta-evaluations have

not been done in the domain of SWS discovery so far except for the work by Tsetsos et al. [17], which is the one most closely related to ours. We share the author's opinion that there is a lack of established evaluation metrics, methodologies and service test collections and agree with them that further analysis is needed to understand whether and how well-known metrics like precision and recall can be applied to service discovery. Tsetsos et al., however, focus on the weaknesses of coarse-grained binary relevance judgements and suggest to use multi-valued relevance judgements instead to exploit the fact that most service matchmakers support different degrees of match instead of a binary match/fail decisions. In contrast, we provided an in-depth discussion why the Cranfield paradigm is not applicable well to SWS discovery evaluation and presented a comprehensive survey and discussion of current service discovery evaluation efforts.

5 Directions for Future Work

5.1 Making Existing Evaluations More Transparent

We found that the existing evaluations generally lack a formal underpinning and fail to clearly discuss the intention behind their design. This makes an objective comparison difficult. As a result of our analysis in Section 3 we derive a preliminary set of questions which should be answered by any evaluation approach.

Assumptions of the test-bed or evaluation. Every evaluation should explicitly state these in order to make its results comparable:

- What are the assumptions on the formalisms / logical language used?
- What is the scope for discovery? E.g. is discovery only concerned with static descriptions or does it also involve dynamic communication?
- What is the expected outcome of the discovery? Are ranked results expected or a boolean match/nonmatch decision? If measures similar to recall or precision are used, are they defined in a meaningful way?

Dimensions measured: Evaluations should clearly indicate and motivate their scope like:

- Runtime performance, such as response time, throughput, etc.
- Scalability in terms of services.
- Complexity of descriptions required.
- Level of guarantees provided by a match. Does it assume manual post processing or support complete automation such that services can be directly executed as a result of the discovery process?
- Standard compliance and reusability. E.g. can existing ontologies be reused?

Probably because of the complexity of the matter the existing approaches only address some of the points above. We believe by providing this initial list of criteria we work towards making evaluations more transparent. This catalog might help to classify test beds and make it easier to find a suitable candidate for

a planned evaluation. In addition, by answering the questions above explicitly, the designer of a test set will increase the actual value of it. This is particularly true since it helps to obtain a more objective result, given that current test sets are mainly created after a particular solution has been developed and might be biased towards that particular solution.

5.2 Towards a Standard Evaluation Methodology and Test Bed

None of the current evaluation approaches provides a comprehensive discussion of the theoretical foundation of SWS discovery evaluation even though such a discussion is necessary to justify the design decisions made by any evaluation approach and ultimately to agree upon a standard way of evaluation.

This paper provides the first comprehensive summary of the current state of the art in this field. As such it hopefully serves as an important first step towards s standard evaluation methodology and test bed for semantic service discovery. Only such an agreed-upon standard will allow to effectively compare approaches and results in an objective way, thereby promoting the advancement of the whole field as such. On the way to this standard, we identify the following rough roadmap for future work.

1. The set of possible dimensions of evaluation have to be clearly identified and motivated (*what to evaluate*).
2. For each of these dimensions suitable means of measurement have to be designed and evaluated (*which criteria to use and how to measure them*).
3. The general requirements to the evaluation process itself have to be identified (*how to achieve validity, realiability and efficiency*).
4. According to these requirements a common semantic service discovery test bed needs to be established, which ultimately allows to effectively evaluate and compare existing solutions with regard to all the dimensions in a unified way. This will clearly be a continuous effort.

6 Summary

We examined the state of the art of evaluation of SWS discovery. We discussed the general applicability of the Cranfield paradigm predominantly used for evaluation in IR and argued that this well-understood paradigm does not map directly to the domain at hand. We continued by presenting an exhaustive survey of the current evaluation approaches in SWS discovery and found that the few existing approaches use very different settings and methodologies highlighting different aspects of SWS discovery evaluation. A thorough discussion of the effects of the decisions in the design of the evaluation on the results of that evaluation is missing so far. We hope that this paper serves as a starting point towards a more systematic approach to SWS discovery evaluation and provided suggestions for future work in this direction.

References

1. McIlraith, S.A., Son, T.C., Zeng, H.: Semantic web services. IEEE Intelligent Systems **16** (2001) 46–53
2. Tichy, W.F., Lukowicz, P., Prechelt, L., Heinz, E.A.: Experimental evaluation in computer science: a quantitative study. Journal of Systems and Software **28** (1995)
3. Tichy, W.F.: Should computer scientists experiment more? IEEE Computer **31** (1998) 32–40
4. Harman, D.: Overview of the first Text REtrieval Conference (TREC-1). In: Proceedings of TREC-1, Gaithersbury, Maryland, USA (1992)
5. Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum (CLEF 2001), Darmstadt, Germany (2001) 355–370
6. Khalid, M.A., Fries, B., Kapahnke, P.: OWLS-TC - OWL-S service retrieval test collection version 2.1 user manual (2006)
7. Klusch, M., Fries, B., Sycara, K.: Automated semantic web service discovery with OWLS-MX. In: Proceedings of the 5th Intern. Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan (2006)
8. Petrie, C., Margaria, T., Küster, U., Lausen, H., Zaremba, M.: SWS Challenge: status, perspectives and lessons learned so far. In: Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS2007), Special Session on Comparative Evaluation of Semantic Web Service Frameworks, Funchal, Madeira-Portugal (2007)
9. Fischer, T.: Entwicklung einer Evaluationsmethodik für Semantic Web Services und Anwendung auf die DIANE Service Descriptions (in German). Master's thesis, IPD, University Karlsruhe (2005)
10. Blake, M.B., Cheung, W., Jaeger, M.C., Wombacher, A.: WSC-06: the web service challenge. In: Proceedings of the Eighth IEEE International Conference on E-Commerce Technology (CEC 2006) and Third IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (EEE 2006), Palo Alto, California, USA (2006)
11. Toma, I., Iqbal, K., Roman, D., Strang, T., Fensel, D., Sapkota, B., Moran, M., Gomez, J.M.: Discovery in grid and web services environments: A survey and evaluation. International Journal on Multiagent and Grid Systems **3** (2007)
12. Åberg, C.: An Evaluation Platform for Semantic Web Technology. PhD thesis, Department of Computer and Information Science, Linköpings Universitet Sweden (2007)
13. Sîrbu, A., Toma, I., Roman, D.: A logic based approach for service discovery with composition support. In: Proceedings of the ECOWS06 Workshop on Emerging Web Services Technology, Zürich, Switzerland (2006)
14. Sîrbu, A.: DIP deliverable D4.14: discovery module prototype (2006)
15. Anonymous: RW2 project deliverable D2.3: prototype implementation of the discovery component (2006)
16. Saracevic, T.: Evaluation of evaluation in information retrieval. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR95), Seattle, Washington, USA (1995) 138–146
17. Tsetsos, V., Anagnostopoulos, C., Hadjiefthymiades, S.: On the evaluation of semantic web service matchmaking systems. In: Proceedings of the 4th IEEE European Conference on Web Services (ECOWS2006), Zürich, Switzerland (2006)