Deployed Semantic Services for the Common User of the Web: A Reality Check

Matthias Klusch and Xiguo Zhing
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3, 66121 Saarbruecken, Germany
klusch@dfki.de

Abstract

Where are all the semantic Web services today? In this paper, we report on quantitative results of searching the surface Web and the prominent citeseer archive as one element of the deep Web for publicly accessible semantic Web service descriptions written in OWL-S, WSML, WSDL-S or SAWSDL by means of the specialized meta-search engine Sousuo 1.6.

1 Introduction

Regarding the impressive efforts, advances, and investments in semantic Web technology during the past decade, the common user of the Web might expect a reasonably fair number of publicly accessible semantic services at her fingertips when searching for them today. However, according to a focussed search for such service descriptions we conducted throughout the year 2007 with a specialized meta-search engine Sousuo this does not seem to be the case. We found not more than about one hundred semantic service descriptions in prominent formats like OWL-S, WSML, WSDL-S and SAWSDL in the surface Web as well as in the scientific archive citeseer. This quantity appears tiny compared to, for example, the more than half million RDF sources indexed by the semantic Web search engine Swoogle, and several hundreds of validated Web service descriptions in the standard format WSDL found by Sousuo 1.6 in the Web.

Of course, one could argue that this low number of public semantic Web services comes at no surprise for two reasons. First, semantic Web service technology with a standard description language, SAWSDL, announced just recently (in August 2007), is still considered immature by both academia and industry which provides insufficient common ground supporting its exploitation by potential end users. Independent from this (WSDL became official recommendation of the Web Consortium only few months before SAWSDL, in June 2007), one could have expected the massive research and development activities of the community around the globe in the past ten years to have produced a considerable amount of publicly visible semantic Web service descriptions that are not merely stored in internal project repositories.

Second, one might argue that it is not clear whether the surface Web and academic publications are the right place to look for semantic Web services, as many of them would be intended for internal or inter-enterprise use but not visible for the public. Though this is one possible reason of the low quantities reported above, there is no experimental evidence in favor of, or against this claim yet. In any case, this merely quantitative result might motivate the community to invest more into raising the public awareness and taking up of developed semantic business service applications by the common Web user.

In our initial search experiment we were particularly interested in the following: How many semantic Web service descriptions are actually accessible to everyone searching the Web for them? How many of these descriptions are written in the standard format SAWSDL, and the non-standard but prominent ones like OWL-S and WSML? What is the distribution of their geographic locations and application domains? How many of these semantic service descriptions are syntactically valid? Finally, how many links to semantic Web service descriptions can only be found in academic publications such as those in the prominent scientific archive citeseer as one element of the deep Web?

In the remainder of this paper, we summarize our quantitative experimental results. [1] In section 2, we

[1] Please note that in this first experiment, we searched for semantic Web service descriptions independent from whether they

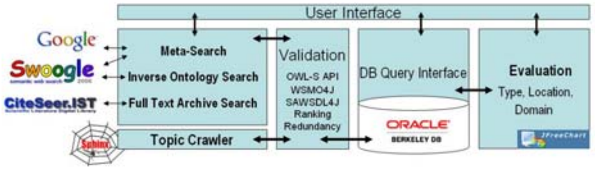

Figure 1. General architecture of Sousuo.

briefly describe the meta-search engine Sousuo 1.6 with which the initial search has been performed together with its testing environment. The performance of both the search engine and its topic crawler followed by the search results are provided in section 4, while we conclude our report in section 5.

2 Meta-search engine Sousuo

The purpose of the specialized meta-search engine Sousuo for semantic Web services is to search the surface Web and the scientific archive citeseer for semantic Web service descriptions in OWL-S, WSML, WSDL-S and SAWSDL.

2.1 Architecture

The general architecture of Sousuo is shown in figure 1.

Overview. Users may select any combination of the following search methods of Sousuo to search for semantic Web services.

- Meta-search (MS) through most prominent search engines Google with A9,
- Sousuo's own focussed topic crawler (TC) based on the WebSphinx crawler,
- Inverse ontology based search (IOS) via Swoogle, and
- Full text scientific archive search (FTAS).

Sousuo considers returned links to syntactically valid service descriptions relevant. For validation, Sousuo uses the validators of publicly available OWL-S API, WSMO4J API, and SAWSDL4J API. Depending on the validation result, it determines the relevance ranking score of each link, that is a score of 1.0 if the service description is syntactically valid, 0.5 if validation failed due to minor syntax errors, and 0 else. Validation is complemented by checking whether the link has been already stored in the local database (Open Berkeley XML database[4]).

This database can be queried and evaluated by the user according to the distribution of geographic locations, description formats, domains and categories of semantic Web services, as well as the coverage of the total result returned for any combination of the different search methods. Sousuo 1.4 also informs about the actual performance of both its topic crawler and the whole meta-search engine. Sousuo has been implemented in Java and is publicly available through the software portal semwebcentral.org dedicated to semantic Web software.

Meta-search and topic crawler. The meta-search of Sousuo is restricted to querying Google, Swoogle, and A9 through their API with predefined and user given search keys. Predefined search keys focus on links to files of type, for example, filetype:{wsdl sawsdl, wsml wsml, owls service, owls profile, owl jp, owl kr, owl tw, owl cn, owl service Profile, owl profile, wsdl annotation }. This is complemented by a simple focussed topic crawler which performs a recursive depth-first search taking the continuously growing set of (initially given) links in the local database as root and base set, and terminates with a time-out or given maximum of search depth reached. Validation, ranking, and redundancy checking of found links are as described above.

Inverse ontology based search. This search method is looking for services that reuse ontologies imported by services that have been already located by Sousuo. For this purpose, Sousuo first checks for each link to a service stored in the database the assigned set of individual ontologies required to understand its semantics. It then searches for service descriptions that contain one or multiple of these selected ontology links through respective queries to Swoogle, A9 and Google. The intuition behind this inverse ontology based search is that semantic Web services share ontologies but may not be fully indexed.

Full text archive search. Sousuo queries the scientific archive citeseer via its API for references to papers on semantic Web services, retrieves the respective documents in the answer set formatted in pdf, and scans each of them for embedded relevant links to SWS descriptions. These links then get validated, ranked,

are grounded in deployed Web services in WSDL, or not. In fact, a comparative analysis of world widely deployed Web services and semantic Web services is out of the scope of this paper. We are currently working on Sousuo 2.0 to support such further experiments.

checked for redundancy, and stored in the local database by Sousuo as mentioned above.

Querying the local database. Finally, Sousuo allows the user to search for relevant semantic Web services in its local database, open Berkely XML of Oracle, by means of full text and structured XML based search. SOusuo is available as open source at http://projects.semwebcentral.org/projects/sousuo/

2.2 Testing environment

Search period and hardware. We ran our search experiment from January 1, 2007, to July 25, 2007, November 2007, and January to February 2008, executing Sousuo once every two weeks over 24 hours on a notebook with CPU Intel Core 2 Duo 2.0GHZ, 2GB RAM, and LAN 10 Mbit/s access to the Internet.

Search space and index. Sousuo does not search any surface Web directory such as yahoo. In fact, our respective pre-experiments showed that the resulting answer sets returned by the directories were already covered by the one produced by Sousuo 1.6. However, we are currently working to enlarge the search space of Sousuo version 2.0 by incorporating specialized search engines that claim to provide access to parts of the deep Web such as clusty, intute, and infomine with an open, non-commercial API for inquiries.

The search space of Sousuo 1.6 is the union of the indices of queried search engines, the discovered realm of the focussed topic crawler plus the index of the scientific archive citeseer. This size is, in general, impossible to determine accurately due to the privacy of information on size, redundancy, and coverage. However, the size of the index of Google is estimated with 11.3 billion links[2], while the size of the Swoogle index and citeseer archive is estimated with 1.7 million [3], and 767,558 links, respectively. Besides, our focussed topic crawler explored 11.2 million links in total during the experiment. With an admittedly speculative 90% of an overlap of the latter indices with the Google index, the search space of Sousuo might be estimated with 13 billion pages. The current size of Sousuo's index equals that of its total answer set for the whole search experiment which amounts to 1439 non-redundant, validated links stored in the local database of Sousuo.

3 Performance

As the real relevance set of semantic Web services in the Web is unknown, and impossible to deduce from neither the set of crawled pages nor the answer sets of particular sources queried, we approximate the classical performance measures of precision and recall for Sousuo's topic crawler by means of the so called target recall and target precision as defined in figure 2 according to [5].

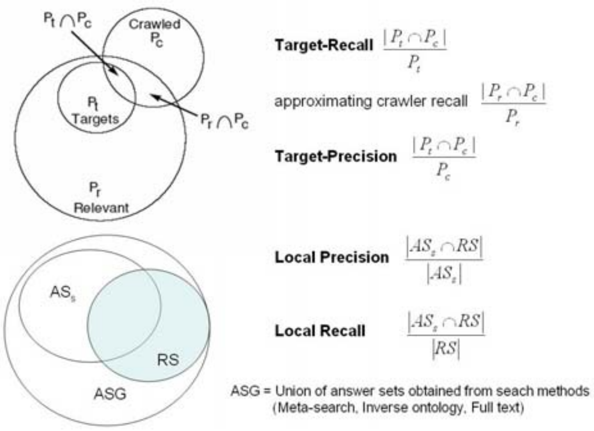
$P_t \cap P_c$
Crawled
P_c
P_t
Targets
$P_r \cap P_c$
P_r
Relevant
Target-Recall $\frac{|P_t \cap P_c|}{P_t}$
approximating crawler recall $\frac{|P_r \cap P_c|}{P_r}$
Target-Precision $\frac{|P_t \cap P_c|}{P_c}$
AS_s
RS
ASG
Local Precision $\frac{|AS_s \cap RS|}{|AS_s|}$
Local Recall $\frac{|AS_s \cap RS|}{|RS|}$
ASG = Union of answer sets obtained from seach methods
(Meta-search, Inverse ontology, Full text)

Figure 2. Target precision and recall.

3.1 Performance of focussed topic crawler

Figure 3 shows the average target precision and recall of the focussed topic crawler of Sousuo. It explored around 11.2 million links in total with fairly reasonable throughput of 46 links per minute during the experiment. Regarding its focussed search the target precision is comparably fair enough as well [5].

3.2 Performance of Sousuo

For measuring the precision of the search enginge Sousuo, we determine the classical ratio between the size of the intersection of its answer set AS with the relevance set RS and the size of AS. The answer set of Sousuo equals the set of valid links taken from those returned by its topic crawler, the search using Google and A9 API which answer set is limited to 1k, respectively, 10k links per day, the inverse ontology based search via Swoogle, and the full text search through citeseer. The relevance set, however, is restricted to the subset of (manually determined) relevant links of the total union of answer sets produced during the search period.

[2] See http://www.linksandlaw.de/news234-indexgroesse.htm

[3] See http://swoogle.umbc.edu/index.php?option=com-swoogle-stats

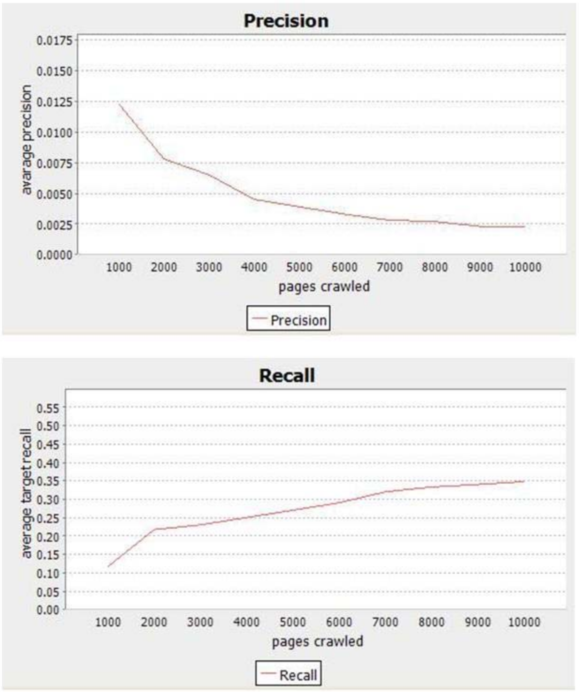

Figure 3. Target precision and recall of Sousuo's focussed topic crawler.

Figure 4 displays the fairly high average precision and recall of Sousuo while its target precision and recall is shown in figure 5.

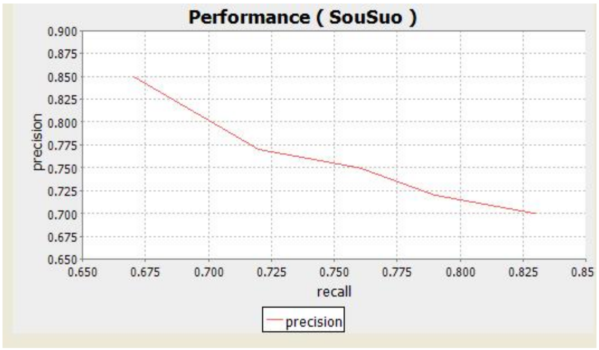

Figure 4. Local precision and recall of Sousuo.

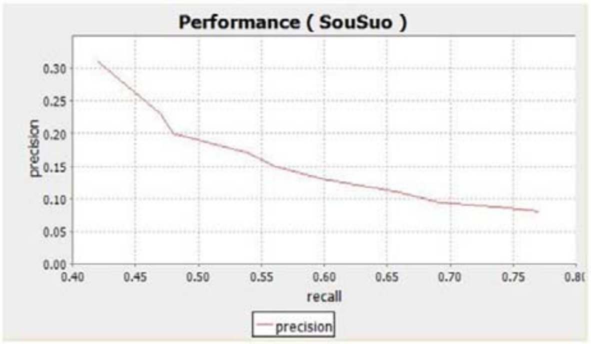

Figure 5. Target precision and recall of Sousuo.

4 Experimental results

The total number of semantic Web services in OWL-S, WSML, WSDL-S and SWASDL including the test collections amounts to 1439 of which only about four percent (65 services) are available outside these collections in the surface Web and through citeseer. Figure 6 shows the number of relevant links found by each of the individual search methods of Sousuo without redundancy checking and test collections OWLS-TC2 [2] and SWS-TC [3]. These quantities did not change as a result of our final search run in December (67 semantic Web services, and 600 non-redundant WSDL services found in December).

	Meta-Search	Citeseer	TC	IOS
Meta-Search	35	4	6	10
Citeseer	4	11	4	2
TC	6	4	29	8
IOS	10	2	8	20

Table 1: Answer sets with links to semantic Web service descriptions (in OWL-S, WSML, WSDL-S and SAWSDL) returned by Sousuo for individual search methods.

The overlapping of answer sets from individual search methods of Sousuo for those services not included in the above mentioned test collections is summarized in table 1. It shows, for example, that the full text archive search returned valid links to semantic Web services in the archive citeseer that were not returned by Google and A9, and a few that have not been returned by any other method. Despite its functional simplicity, the same result holds with our focussed topic crawler (TC). This is in contrast to the inverse ontology based search (IOS) which answer set is completely covered by those of the other search methods.

4.1 Support of semantic Web service formats

Figure 7 shows the quantitative distribution of semantic Web service descriptions in prominent formats,

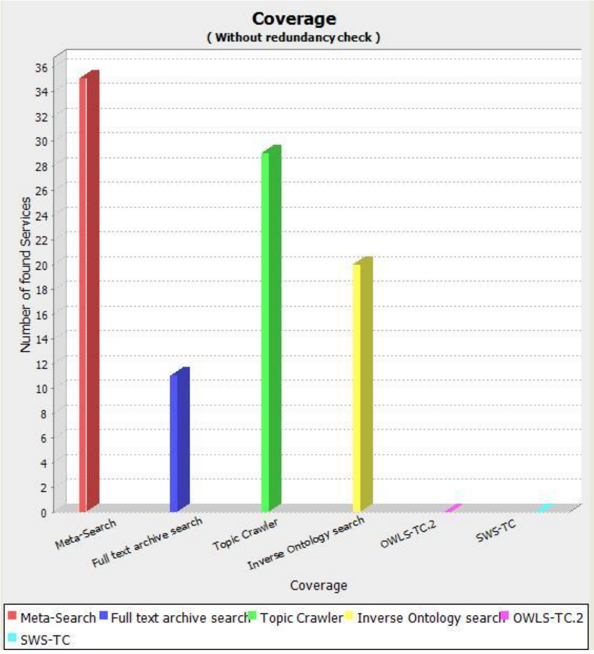

Figure 6. Number of relevant links found by individual search methods of Sousuo.

that are OWL-S, WSML, WSDL-S and SAWSDL, without counting those in the test collections, while figure 8 shows its relation to the number of found Web service descriptions in WSDL (.wsdl) that are syntactically valid.

Given the historic evolution of semantic Web service technology, and its reasonably fair software support today, it comes at no surprise that the quantities of semantic service descriptions in OWL-S still outnumber the considered alternatives. Remarkably, there are less public WSML services than for WSDL-S and SAWSDL together. Though SAWSDL became a proposed recommendation by the Web Consortium just recently, its software support world wide still appears comparatively negligible - which might rapidly change in near term in response to its standardisation.

Apart from two rather medium sized semantic Web service retrieval test collections for OWL-S, that are the OWLS-TC2 [2] and the SWS-TC[3], we did not find any other retrieval test collection in the Web. Please note that a mere collection of services such as the one available for SAWSDL at http://www.w3.org/2002/ws/sawsdl/ is not sufficient for this purpose, since it would have to be extended with set of queries and respective relevance sets.

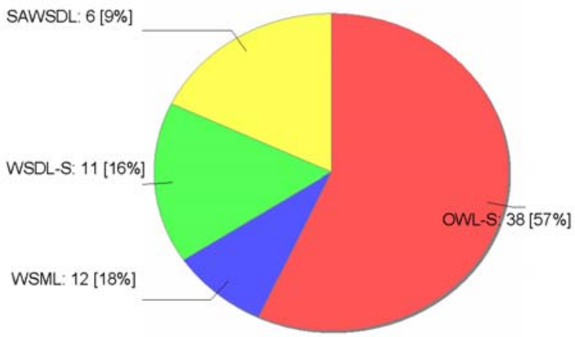

Figure 7. Distribution of semantic Web service formats OWL-S, WSML, WSDL-S and the standard SAWSDL.

4.2 Geographic distribution

Figures 9 and 10 provide an overview of the geographic distribution and locations of the publicly accessible semantic Web services. Regarding the history of the semantic Web vision, in particular the early joint start between researchers in the US and the EU on DAML+OIL, OWL and OWL-S, as well as the massive funding of WSMO related projects in this field by the European Commission, it might not come at a surprise that, according to the quantities reported, the domain appears clearly dominated by the US and Europe while being remarkably close to each other.

However, what surprised us most is that, though major projects in the area exist, we could not find any valid semantic Web service description published in the rest of the world publicly, in particular the Asia and Pacific rim. Additional personal communication with few selected research groups at universities in these regions revealed that, if semantic Web service descriptions do exist at their site, the public retrieval from specific project related repositories is prohibited, hence invisible to any search engine. In general, we appreciate any reference by the interested reader to publicly accessible semantic Web services in one of the considered formats, in particular if published in the above mentioned geographic regions.

4.3 Internet domains and business categories

Figures 11 and 12 show the distribution of the domain and business category of found services outside

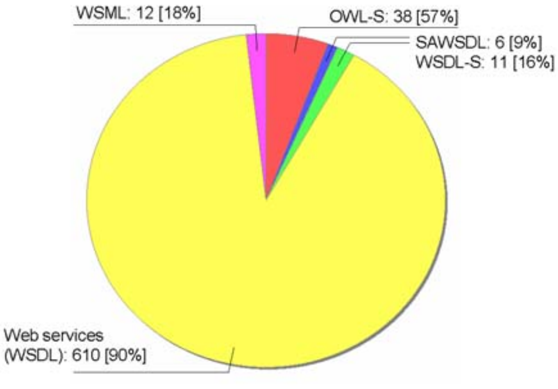

Figure 8. Support of Web services in WSDL Vs. semantic Web services.

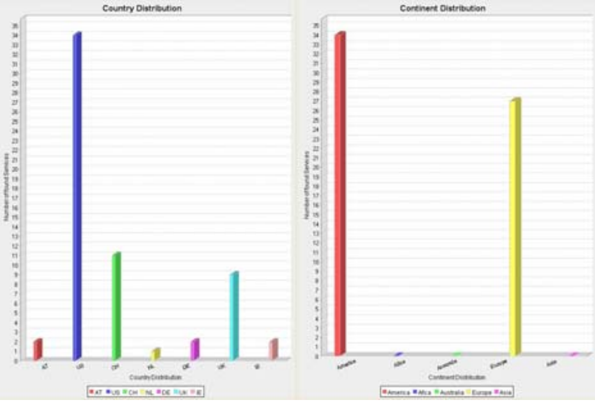

Figure 9. Geographic distribution of semantic Web services.

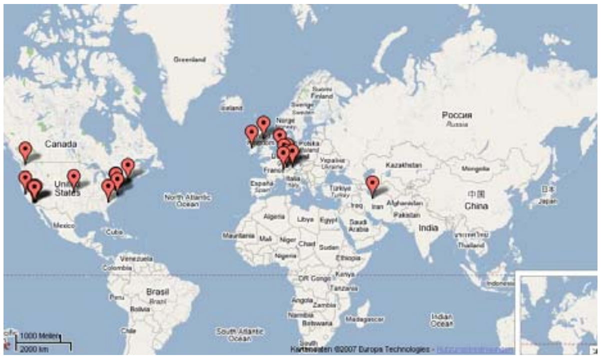

Figure 10. Geographic locations of semantic Web service providers.

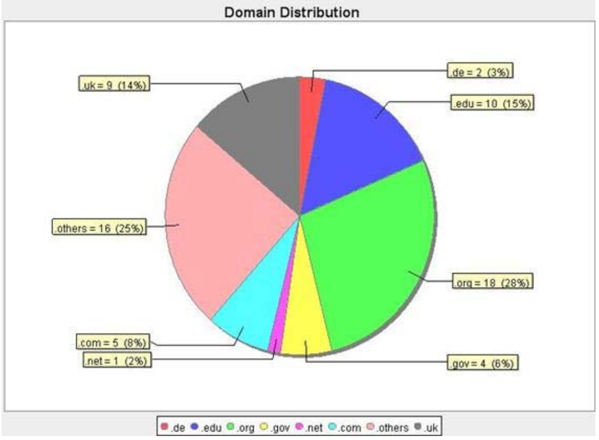

Figure 11. Internet domains of semantic Web services.

the test collections. Business and travel are the most common categories, followed by finance and education.

In compliance with the prevalent geographic distribution and location of semantic Web service links in the US and the EU, the majority of links from Internet domains devoted to commerce (.com, 8 %), organisational (.org, 28%) and educational institutions (.edu, 15%) is hosted in these world regions. Remarkably, most of these semantic Web services publicly accessible in the EU are located in the UK.

The most common business domain for semantic Web services according to their naming or statement in the profiles are business (16%) and travel (17%), followed by education (11%), finance (11%), and government (6%). One third of found semantic Web service links, however, belongs to a variety of other domains of smaller size such as sports and health.

5 Conclusions

The preliminary results of our experimental searching for semantic Web services in the surface Web by use of a specialized meta-search engine is rather desillusioning. We found not more than around 1500 indexed semantic service descriptions in OWL-S, WSML, WSDL-S or SAWSDL in the Web, of which only about four percent are located outside special test collections like the OWLS-TC2. This quantity appears tiny compared with the sheer volumes of estimated thirty billion and one million indexed resources in the Web, respectively, semantic Web encoded in RDF and OWL.

As mentioned above, the reported preliminary experimental result does not reflect the strong research efforts carried out in the SWS domain world wide in

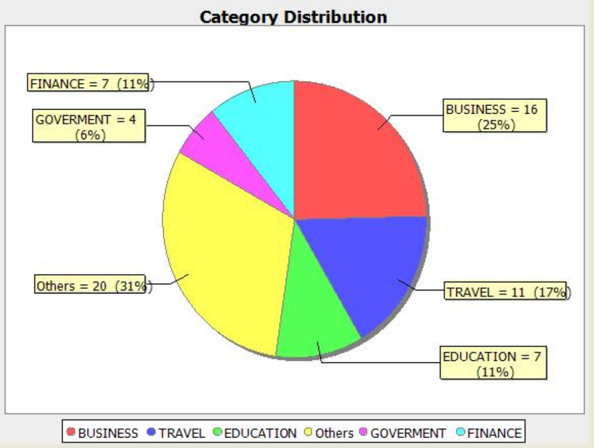

Figure 12. Business categories of semantic Web services.

the past few years, independent from the status of maturity of SWS technology and implied low adoption by end users yet. Nevertheless, the result might encourage the community as a whole to increase its visibility and awareness to the common Web user outside the community and savvy project teams also by publishing a significant number of SWS show cases in the surface Web.

Although one could have expected these results, in particular the majority of semantic Web services being published in protected internal project repositories and other sites of the deep Web[6], there was no experimental evidence available in favor of this claim or against. On the other hand, there still is plenty of space left to search both the surface and the deep Web: The search space of Sousuo in its current version is limited to only few selected indices of freely accessible and prominent search engines with open API, that are Google, A9, Swoogle, and the scientific archive citeseer.

However, raising awareness by a significant number of show cases seems senseless if it is not complemented by community efforts to equip end users with easy to use software support for building, sharing and reusing semantic Web services. Unfortunately, this is support is missing either, despite the variety of SWS related software available at relevant open source software portals such as semwebcentral.org and sourceforge.net. Though, it remains to be seen whether and to what extent these tools or the experience of developing them can be exploited to support the standard SAWSDL for which, apart from the project internal MWSDI infrastructure and Lumina search engine, no public software support exists yet. Sousuo is available at http://projects.semwebcentral.org/projects/sousuo/ Ongoing work is on improving the search methods of Sousuo for an updated version 2.0.

References

[1] L. Ding, T. Finin, A. Joshi, R.S. Cost, J. Sachs: Swoogle: A Semantic Web search engine and metadata engine. Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, 2004.

[2] OWLS-TC2:
http://projects.semwebcentral.org/projects/owls-tc/.

[3] SWS-TC:
http://projects.semwebcentral.org/projects/sws-tc

[4] Oracle's Open Berkeley XML Database:
http://www.oracle.com/database/berkeley-db/xml/index.html

[5] G. Pant, P. Srinivasan, F. Menczer: Crawling the Web. In M. Levene and A. Poulovassilis, eds.: Web Dynamics, Springer, 2004.

[6] C. Sherman, G. Price: The Invisible Web: Uncovering Information Sources Search Engines Can't See. Cyberage Books, 2001. The Deep Web: http://en.wikipedia.org/wiki/Deep-web.

[7] SouSuo 1.3:
http//:projects.semwebcentral.org/projects/sousuo/.

[8] Swoogle:
http://swoogle.umbc.edu/