# Semantic Web evolution: towards the usability of data

Adam Westerski
Universidad Politecnica de Madrid
Madrid,Spain
westerski@dit.upm.es

**In the following paper we describe the evolution of the Semantic Web area throughout the years, with a special focus on the most recent years. We present how the picture of the Semantic Web has gradually changed from a very idealistic vision to a more realistic scenario. The assumption shall be backed up with both examples from the state of the art in the Semantic Web research and commercialization attempts.**

## I. Introduction

The Semantic Web vision has been present in the IT world for quite a lot of time [1]. Although many foretold brilliant results and great success, it never came to be an important part of the Internet. Throughout its existence there have been many attempts to apply Semantic Web in different domains (i.e. medical science, general formalization of knowledge). It was believed that penetration of single domains of industry could bring the ultimate success. The metadata produced within one domain would grow and eventually connect through various concepts with other domains. Some projects did indeed succeed in research laboratories and on paper in terms evaluation. Nevertheless, all failed to make the final step and become mainstream or at least a significant part of infrastructures used by common people in a given area.

The creation of corner stones of the Semantic Web technologies such as RDF [2] was followed by the first visions of the Semantic Web [1, 3]. The original publications did not assume the Semantic Web to be a form of artificial intelligence on the Web [3]. Nevertheless, in practice the catchy term of the intelligent web led to exploration of metadata in terms of sophisticated reasoning techniques. One of the most notable examples is the topic of the Semantic Web Services. This domain has been present almost since the birth of the Semantic Web idea and therefore it has been vastly explored since. The various research attempts [4][5] were envisioned to succeed where SOAP based services failed to reach mass usage like initially intended. In practice, the research is still ongoing but the hype and anticipation have long passed. With time, the original tools and theoretical frameworks became too complex to introduce in practice (which most notably was also one of the critical problems of the SOAP services). Similar patterns could be pointed in many other areas. However, it has to be noted that valuable lessons have been learned and the way of thinking and directions of research have changed since towards greater simplification (i.e. in terms of Semantic Web Services, WSMO-Lite [6] ).

The following paper does not aim to present any breakthrough conclusions about the state of the Semantic Web, nor does it try to envision its future. The value of this work is strictly informative and is supposed to show a point of view on Semantic Web through the eyes of one of many researchers involved in this community. While not being the first [7], this paper takes an attempt to point of some trends in the contemporary Semantic Web community and the changes with respect to previous ways of thinking. In the following sections we shall present the rise and the evolution of the Linked Data initiative – currently the most prominent branch of the Semantic Web (see Sec. 2). Next, we shall highlight some of the problems that the research community faces in terms of data utilization and putting it into practical use (see Sec. 3). Finally, we shall conclude with the description of contemporary attempts and ideas how to bring the Web and the Semantic Web into synergy (see Sec. 4). As a supplement we also present an additional section with interesting new research ideas and emerging trends that are not yet fully evaluated (see Sec. 5).

## II. Semantic Web and Linking Open Data

Although the principles of Linked Data were pointed out by Tim Berners-Lee in 2006 [8] it was for the Linking Open Data project [9] to make this term rise in popularity. Currently the Linked Data is considered to be the next step in the evolution of Semantic Web idea. It is not what the initial trends envisioned for Semantic Web to become, nevertheless Linked Data is where Semantic Web is at the moment.

There are many mature Semantic Web projects or even standards already published. All of those were supposed to build the foundations of the new Internet but they are not popular neither widely accepted. Some point out the overwhelming complexity as reason for such situation, while others remark small and unrealistic appliances of what has been released so far. The originally envisioned and hyped Artificial Intelligence for the Web has not arrived therefore a simpler solution is taken into account as the first step on the road to reaching the Intelligent Internet. This first step is called Linking Open Data. One of the key concepts of this trend is to provide not only annotations for the Web but also means to interconnect the annotated resources. Amongs others the community maintained webpage provides a constantly updated map of the current Linked Data sets available over the Web (see Fig. 1).

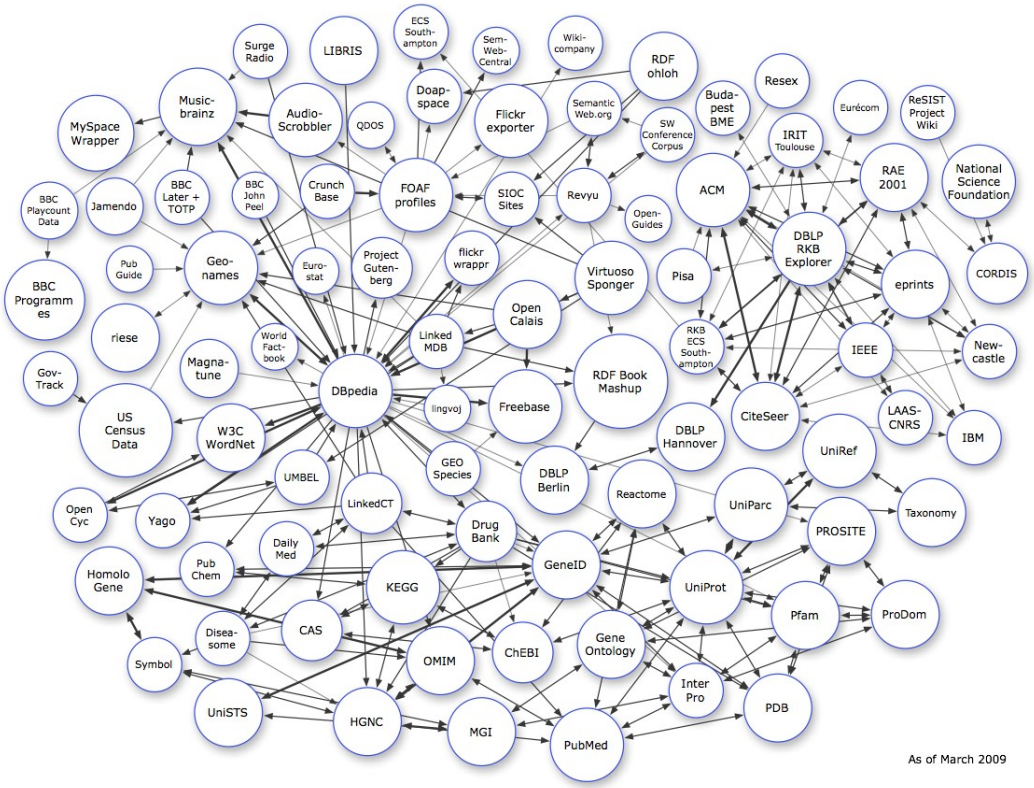


Figure 1. Linked Data Map [9]

The outcome of initial Semantic Web activity is a situation where the metadata and annotations of web resources are already out in the Internet but data is still not interlinked and therefore hard to use in practice. Selected websites or particular groups of interest have their web resources annotated but outside of those communities the information is not usable or no one has any interest of using it. Therefore, the Linking Open Data initiative is also about reaching out to the people - the Internet community. Not some selected, enclosed groups like before but more massive audience. The goal is to introduce easy to use paradigms for annotation of the web content [10][11] and functional applications that take advantage of those annotations (semantic search engines [12]13][14][15], Semantic Web based portals like Twine[16]).

By some the term Web 3.0 is attached to the state of Internet where publishing the Linked Data becomes a norm, not a hyped buzzword among researchers. This is because it harvests on the idea that was the corner stone of Web 2.0 success - social networking. A great number of the main stream activities within Linked Data put big emphasis on Internet users, user activities and virtual societies in the modern Internet. There is a lot of discussion and research done about annotations in collaborative environments that dominate the contemporary Internet environment.

In the end, the ultimate goal remains the same for the Linked Data. Similarly as the aforementioned Semantic Web, is supposed to deliver more user friendly Internet through appliance of machine understandable content and mediation between human and machine.

## III. Towards usability of Data

### A. *Community engagement*

Community engagement is arguably one of the biggest failures of the early Semantic Web. Most of the work done in the area assumed that at some point of time the Web will evolve by itself and the information produced by users will combine into one big graph of the World Wide Web. Unfortunately, this never happened. With time, it has been realized that people shall not deliver data sets undermining their web content unless they can receive immediate benefits from production of their data in computer readable formats.

The activities around the aforementioned Linked Data initiative (see Sec. 2) try to handle this problem. The approach is to encourage institutions and organizations to let the data undermining their applications to be available to public. Partially (among others due to to massive involvement of the W3C leader Tim Berners-Lee), this is becoming successful. Most prominent examples beging such enterprises as Reuters [17] or New York Times [18]. Following those examples other companies publish their data available in different standards and share them free or based on subscription fees. At this moment it is hard to determine

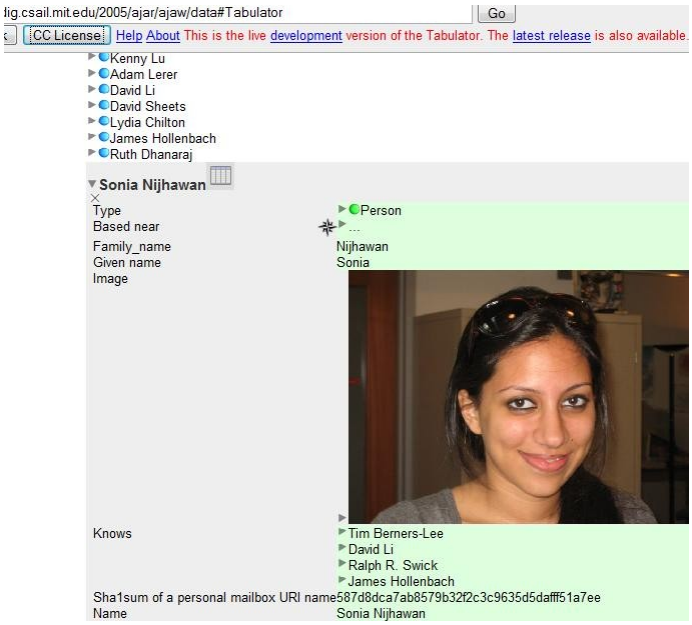


Figure 2. Sample of interface and navigation for a Semantic Web based application – the Tabulator [23].

how successful this initiative is or will prove in the future. One is certain – the Linked Data cloud is growing.

Although the progress is unarguable, a number of problems still remain. Most notably, the casual user community still does not seem to notice the need to publish data or the benefits that can be derived from such activities. Without the everyday user contributing their part to the Linked Data cloud the amount of metadata may never grow to a satisfactory size. According to research made by some companies the user created content is already a significant part of the overall data published on the Internet [19][20]. Additionally according to Internet analysts this number shall grow further largely suppressing Internet content as it was known in the XX century and early years of XXI century. In practice, this means that support of the community is very valuable and might become the key factor to realize Linked Data vision and later Semantic Web.

Fortunately in terms of technology, just like in many other areas, the foundations for publishing metadata about community sources have been already created [21]. The authors of the SIOC initiative even try to engage the community by organization of various contests for metadata use cases [22]. Nevertheless, the effective way to encourage the massive production of metadata for community created content is still to be found.

### *B. Interfaces for Semantic Web applications*

The publishing of the data across the web is only part of the success. One has to remember that all this information should be put into use. Unless there is a consumer of the data and some realistic benefits there is no real need to publish it no matter how brilliant the research theories behind.

One of the approaches is to design interfaces that allow to browse all the semantic data behind the Internet resources. This exploits the graph nature of RDF and in general the vast amount of links between the data. Normally its not as easy to notice all the connections while with Semantic Web data, due to notation standards, it comes natural. Among the most notable projects developed for the end user Tabulator [23] and Disco [24] can be mentioned. However, in practice, those tools have so far proven their value only for developers and researchers who can quite easy analyze the Semantic Web data and construct applications (see Fig. 2).

### *C. Road from complex to simple applications*

A different approach, opposed to construction of user interfaces to Semantic Web data is the usage of this information in the backend. The goal is to enhance current user experience without the user to know about the mechanics below. The users do not need to be aware that they produce or consume machine understandable metadata, only perceive the benefits and improved user experience they are granted through their actions.

Unfortunately, although this approach might seem very accurate, it has proven to be quite difficult to realize. Applications like Tabulator or other Semantic Web data browsers require too much specialistic knowledge to be directed towards an average Internet user. On the other hand, even domain specialists that are not familiar with Semantic Web (or even IT) find it often too complex to construct domain specific ontologies to even deploy their data (with tools like for instance Protege [25]).

Partially due to the success of Web 2.0 and very simple concepts like *tags* gaining popularity, the Semantic Web community has also noticed the need to introduce more straightforward solutions [7]. As a result we can see various attempts to merge the Semantic Web with the current Internet mainstream on both application and data production level.

## IV. Assimilation of the Semantic Web to the current Internet

With the Semantic Web tools and ideas going mature and not brining able to cause the final switch of the Internet into the envisioned Web of Data, many research efforts have been directed toward more subtle methods of introducing the key concepts once laid down by Tim Berners-Lee [1].

### A. Research focus

One of the important trends in the area is the research on easy inclusion of machine readable metadata in the normal HTML pages. The initial solutions due to lack of proper research used various HTML tags to link the RDF documents (i.e. SIOC ontology [26] metadata is connected to the page it described through a *link* tag). However, as mentioned before, creation of metadata and the understanding of Semantic Web technologies never got wide acceptance of the Internet community. Therefore, researchers started to search for easier and more straightforward ways to describe HTML content.

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/">
   <h2 property="dc:title">The trouble with Bob</h2>
   <h3 property="dc:creator">Alice</h3>
   ...
</div>
```

Figure 3. Annotations included in the XHTML file. Example of RDFa [27].

One of the attempts is called RDFa [27] - an extension of XHTML that defines principles to create page annotations with XHTML tags. It does not relieve of the complexity of RDF because the principals of the framework are still kept. However, one important step is made, RDF files are no longer linked as separate entities, the annotations and the HTML content is placed in one physical file (see Fig. 3).

A slightly different approach is presented by the Semantic Sitemaps [11]. Being an extension of the Sitemap protocol [28] the Semantic Sitemaps try to standardize the way datasets are connected to Internet resources. This is specially meant for large datasets that do not explicitly connect to the exact content presented on the website. For instance, a Semantic Sitemaps protocol is used to describe ways to access metadata behind the DBpedia [29] (the SPARQL endpoint, the RDF dump or direct access to single resources).

Also recently, an interesting initiative has been proposed by a consortium of various Semantic Web and Web institutions – Common Tag.

Apart of the aforementioned initiatives to produce RDF originated data within the webpage, a slightly different approach has also gained a lot of interest – the microformats [30]. Although, the basic concept has been know for quite long, with the birth of RDFa and Common Tag, microformats get a lot of attention yet again. In practice, the idea behind is very similar, through a number of XHTML tags and custom classes annotate the web conent. The differences are hidden within the technical realization. The microformats were not intended to expose links between the data while RDFa and Common Tags are designed with Linked Data in mind from the very start.

### B. Commercial attempts

One of the most important and interesting topics, regarding the assimilation of Semantic Web to the current Internet, is the interest of industry and, as a consequence, release of products that utilize Semantic Web technologies. Up to a small degree this was present for a long time. However the recent years how shown an exceptional rise of activity in the area. The results of increasing Semantic Web interest can be seen within small startups as well as long established Web companies.

Within the startups, one of the most recognizable companies up to date is the Radar Networks and their flagship product called Twine [16]. The path of their choice is to construct a social portal and use Semantic Web technologies and ideas for better data management and processing in the backend. Although, not entirely successful Twine is a good example of a model where user is completely unaware of the complexity behind the Semantic Web. Through a standard interface the user creates the semantically linked data and annotates the content. Later on, the benefits can be seen in forms of links between the content that all users have created.

On the side of large companies there has also been a lot of interest in metadata publish across the web. However, the Web search companies such are Yahoo or Google seem to be more interested so far in the data consuming side rather then production. Nevertheless, such activity also helps to increase the popularity of the Semantic Web among casual Internet users. By providing enhanced capabilities for websites that produce metadata both search engines encourage the production of metadata.

The most active in the area has been Yahoo!. Their engagement in research (not only Semantic Web) and participation in the research community is truly great (i.e. during WWW 2009 conference the company had most accepted paper of any organization [31]). With regard to the Semantic Web technologies, one of their most important contributions is called Search Monkey [32]. The idea behind is different rendering on search result list for webpages with metadata. For instance if products on a particular website are annotated with hReview [33] microformat then Yahoo! will display such results with a graphical representation of their rating next to the text summary usually present for all normal sites. Such custom rendering makes the particular search result more appealing and people are likely to choose it over normal results – clearly an advantage and motivating factor to produce metadata. This effect could be compared to publishing advertisements on Yellow Pages – the bigger and more distinctive from competitors the better. However, with respect to search engines such customization is still free.

Another company that recently showed their interest in the Semantic Web technologies in public is Google.

Although a lot less has been disclosed in comparison to Yahoo!, first announcements have been made about Google planning to create similar rich snippets for search results with metadata [34]. In practice, no live demonstration is available up to date but even a mare announcement that the leading Web company will provide support for online metadata drew a lot of attention.

## V. Other directions in the contemporary Semantic Web research.

Some of the research going on in the area of the Semantic Web does directly fall into the category of being a bridge between the current Internet community and the Semantic Web community. Nevertheless, some of those projects still deliver a significant value and perhaps build the foundations for other research and companies to supply the solutions for end user. Within this section we shall provide a discussion of such research attempts.

### A. Semantic Web Search Engines

The interest in the Semantic Web Search Engines has grown considerable during the recent years. Nevertheless their construction still holds many unsolved problems and faces the never ending task of efficiency increase. Some key challenges in the area are: metadata crawling, efficient indexing, efficient index browsing, scalability, resource ranking, results clustering, and most notably further utilization of the results (as in metadata search engine use cases).

The projects in the area have various approaches. Some aim to achieve best possible scalability at the price of offered query capabilities [12], while others seek to provide prototypes that will enable as much features as possible leaving out scalability issues in the background [13]. Nevertheless what can be seen most cases, where the research has been already done in most directions, is the care for the search engine utilization.

The Semantic Web Search Engines contrary to regular search engines as known today, output result lists only with the metadata meant for machines. In most cases, for a human, the results provided by the Semantic Web Search Engine are not directly valuable. Due to the lack of external interest, most of the research teams try to produce their own use cases based on the capabilities of Semantic Search Engines exposed as a web service. Some attempts are based on publishing the public API to the search engines [35] [36] and giving some samples while others include a presentation of complete solutions based on such technology [37]. However, in the end, while some argue that the Semantic Web Search Engines will show their full capabilities when there is enough useful metadata on the Web, the topic still remains open and the amount of practical use cases for this technology is still little.

### B. Scalability

An interesting branch of the Semantic Web that started to gain increasing interest along with the rise of popularity of Semantic Web Search Engines and large data repositories is technology scalability.

In case of the search engines during their development it occurred that performing operations on metadata an allowing more complex queries then just text search results in many new problems then in regular keyword search engines. Maintenance of the index, indexing algorithms, finding the right balance between capabilities and the speed and many others are the concepts that had to be dealt anew in case of the emerging search engines for the metadata

Additionally the research on scalability issues in the Semantic Web area was also provoked by the rise of interest in production of metadata on large scale (i.e. Dbpedia [38]). Large amounts of data encoded in specific format (i.e. RDF) require new ways of processing and storage to let efficient access and update. One of the leading vendors of solutions that have scalability in mind is the OpenLink [39] company with their product called Virtuoso [40].

### C. Trust management

Although this area research within the Semantic Web is not quite popular yet, the initial works show the necessity to analyze the problem. In a situation when the metadata publication gets popular it is possible many people will produce similar information or information on the identical topic. In order to process if correctly one has to have some kind of criteria about the importance of the information. One of the research attempts [41] in this area suggest the inclusion of provenance information inside the metadata. This way one can judge the credibility of the data based on its origin. Therefore it could be recognized wether the data is a fraud or for example the probability of data being true or false based on the distance from the sources considered as trustworthy.

## VI. Conclusions

The progression in the Semantic Web community is clearly viable. In the time span of recent years a lot has been done and is happening at the moment to bring the entire idea closer to being real. Much more thought is given into appliances, realistic scenarios, building consortiums between research and corporate world to establish standards that would be used in practice. The direction and the way of thinking about the entire topic has defiantly changed. The Semantic Web main stream activities are no longer about complex solutions that hardly anyone outside the domain can understand. With all its effort the Semantic Web community is trying to make the technologies finally leave the research laboratories and take their place in the modern Internet. Although it is still early to say, most activities described in this paper point to a conclusion that the introduction and popularization of ideas connected to the Semantic Web will probably never happen like initially envisioned or hyped. So far the changes seem to slowly and gradually come to live through small portions adopted by the Internet community.

## References


[1] T. Berners-Lee, Semantic Web Road map, 1998, http://www.w3.org/DesignIssues/Semantic.html

[2] O. Lassila , R. R. Swick, Resource Description Framework (RDF) Model and Syntax, http://www.w3.org/TR/WD-rdf-syntax-971002/, 1997

[3] T. Berners-Lee, What the Semantic Web can represent, http://www.w3.org/DesignIssues/RDFnot.html ,1998

[4] D. Roman, U. Keller, H. Lausen, J. d. Bruijn, R. e. Lara, M. Stollberg, A. Polleres, C. Feier and D. Fensel, Web Service Modeling Ontology, 2005

[5] A. Ankolekar, M. Burstein, J. Hobbs, O. Lassila, D. Martin, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, K. Sycara, and H. Zeng. DAML-S: Semantic Markup for Web Services. Presented at International Semantic Web Working Symposium (SWWS), 2001

[6] T. Vitvar, J. Kopecky, M. Zaremba, D. Fensel, WSMO-Lite: Lightweight Descriptions of Services on the Web. In Proceedings of the IEEE European Conference on Web Services, IEEE Computer Society Halle (Saale), Germany, 2007

[7] A. Ankolekar, M. Krotzsch, T. Tran, and D. Vrandecic, "The two cultures: mashing up web 2.0 and the semantic web," in WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM Press, 2007, pp. 825-834.

[8] T. Berners-Lee, Linked Data design issues, 1996, http://www.w3.org/DesignIssues/LinkedData.html

[9] Linking Open Data project page, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/

[10] RDFa Primier, http://www.w3.org/TR/xhtml-rdfa-primer/

[11] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, G. Tummarello, Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In Proceedings of the Proceedings of the 5th European Semantic Web Conference 2008.

[12] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, G. Tummarello, Sindice.com: A document-oriented lookup index for open linked data. In International Journal of Metadata, Semantics and Ontologies, 3(1), 2008.

[13] A. Harth, A. Hogan, J. Umbrich, S. Decker, Building a Semantic Web Search Engine: Challenges and Solutions. In Proceedings of the 3rd XTech Conference, 2008, Dublin, Ireland, 2008.

[14] M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, and E. Motta. Watson: A Gateway for the Semantic Web. Poster session of the European Semantic Web Conference, ESWC 2007.

[15] Y. Qu, G. Cheng, H. Wu, W. Ge, X. Zhang, Seeking Knowledge with Falcons. In Semantic Web Challenge, 2008

[16] Twine, http://www.twine.com/

[17] Open Calais website, http://www.opencalais.com/

[18] NYT to Release Thesaurus and Enter Linked Data Cloud, The New York Times Open blog, 2009, http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/

[19] E. Burns, Clikz, Pew: Nearly 50 MM Americans Create Web Content, 2006, http://www.clickz.com/3609461

[20] P. Verna, eMarketer Digital Intelligence, User-Generated Content: Will Web 2.0 Pay Its Way?, http://www.emarketer.com/Report.aspx?code=emarketer_2000421

[21] J.G. Breslin, S. Decker, "The Future of Social Networks on the Internet: The Need for Semantics ", IEEE Internet Computing magazine, November 200

[22] SIOC Data Competition, http://data.sioc-project.org/

[23] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, D. Sheets, Tabulator: Exploring and Analyzing linked data on the Semantic Web, Procedings of the The 3rd International Semantic Web User Interaction Workshop (SWUI06) workshop, Athens, Georgia, 6 Nov 2006.

[24] Disco - Hyperdata browser homepage, http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[25] Protege homepage, http://protege.stanford.edu/

[26] SIOC ontology specification, 2009, http://rdfs.org/sioc/spec/

[27] B. Adida, M. BirBeck, RDFa Premier, Bridging the Human and Data Webs, 2008, http://www.w3.org/TR/xhtml-rdfa-primer/

[28] Sitemaps XML format, http://www.sitemaps.org/protocol.php

[29] Dbpedia Semantic Sitemap, http://dbpedia.org/sitemap.xml

[30] Microformats homepage, http://microformats.org/

[31] Yahoo! Research, Yahoo! at WWW2009 in Madrid, 2009, http://research.yahoo.com/news/2759

[32] Yahoo! Search Monkey webpage, 2009, http://developer.yahoo.com/searchmonkey/

[33] hReview microformat draft specification, 2009, http://microformats.org/wiki/hreview

[34] Introducing Rich Snippets, Google Webmaster Central Blog, 2008, http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html

[35] A Westerski, A. Iqbal, G. Tummarello, S. Decker Sindice Widgets: Lightweight embedding of Semantic Web capabilities into existing user applications. In Proceedings of the 4th International Workshop on Semantic Web Enabled Software Engineering 2008.

[36] M. d'Aquin. Building Semantic Web Based Applications with Watson. Developers' Track of the WWW 2008 conference, Beijing China

[37] R. Cyganiak, M. Catasta, G. Tummarello, Towards ECSSE: live Web of Data search and integration, Semantic Search Workshop at WWW 2009 conference, Madrid, Spain

[38] Dbpedia website, http://www.dbpedia.org

[39] OpenLink company homepage, http://www.openlinksw.com/

[40] Virtuoso homepage, http://virtuoso.openlinksw.com/

[41] O. Hartig, Provenance Information in the Web of Data, Linked Data workshop at WWW 2009 conference, Madrid, Spain