The Evolution of the Semantic Web

John Cardiff

Social Media Research Group,
Institute of Technology Tallaght, Dublin, Ireland
email John.Cardiff@ittdublin.ie

Abstract — **The Semantic Web offers an exciting promise of a world in which computers and humans can cooperate effectively with a common understanding of the meaning of data. However, in the decade since the term has come into widespread usage, Semantic Web applications have been slow to emerge from the research laboratories. In this paper, we present a brief overview of the Semantic Web vision and the underlying technologies. We describe the advances made in recent years and explain why we believe that Semantic Web technology will be the driving force behind the next generation of Web applications.**

Keywords: Semantic Web, Ontology, Web of Data

I. INTRODUCTION

The World Wide Web (WWW) was invented by Tim Berners Lee in 1989, while he was working at the European Laboratory for Particle Physics (CERN) in Switzerland. It was conceived as a means to allow physicists working in different countries to communicate and to share documentation more efficiently. He wrote the first browser and Web server, allowing hypertext documents to be stored, retrieved and viewed.

The Web added two important services to the internet - it provided a very convenient means for us to retrieve and view information - we can then see the web as a vast document store in which we retrieve documents (web pages) by typing in their address into a web browser. Secondly, it provided a language called HTML, which describes to computers how to display documents written in this language. Documents, or web pages, are accessed by a unique identifier called a Uniform Resource Locator (URL) and are accessed using a Web browser. Within a short space of time, the WWW had become a popular infrastructure for sharing information, and as the volume of information increased its use became increasingly widespread.

Although the web provides the infrastructure for us to publish and retrieve documents, the HTML language defines only the visual characteristics, ie. how the documents are to be presented on a computer screen to the user. It is up to the user who requested the document to interpret the information it contains. This seems counterintuitive, as we normally think of computers as the tools to perform the more complex tasks, making life easier for humans. The problem is that within HTML there is no consideration of the meaning of the document, they are not represented in a way that allows interpretation of their information content by computers.

If computers could interpret the content of a web page, a lot of exciting possibilities would arise. Information could be exchanged between machines, automated processing and integration of data on different sites could occur. Fundamentally, they could improve the ways in which they can retrieve and utilise the information for us because they would have an understanding of what we are interested in. This is where the Semantic Web fits into the picture - today's web (the "syntactic" web) is about documents whereas the semantic web is about "things" - concepts we are interested in (people, places, events etc.), and the relationships between these concepts.

The Semantic Web vision envisages advanced management of the information on the internet, allowing us to pose queries rather than browse documents, to infer new knowledge from existing facts, and to identify inconsistencies. Some of the advantages of achieving this goal include [4]:

- The ability to locate information based on its meaning, eg. knowing when two statements are equivalent, or knowing that a reference to a person in different web pages are referring to the same individual.
- Integrating information across different sources − by creating mappings across application and terminological boundaries we can identify identical or related concepts,
- Improving the way in which information is presented to a user, eg. aggregating information from different sources, removing duplicates, and summarising the data.

While the technologies to enable the development of the Semantic Web were in place from the conception of the web, a seminal article by Tim Berners-Lee, James Hendler and Ora Lassila [1] published in *Scientific American* in 2001 provided the impetus for research and development to commence. The authors described a world in which independent applications could cooperate and share their data in a seamless way to allow the user to achieve a task with minimal intervention. Central to this vision is the ability to "unlock" data that is controlled by different applications and make it available for use by other applications. Much of this data is already available on the

Web, for example we can access our bank statements, our diaries and our photos online. But the data is controlled by proprietary applications. The Semantic Web vision is to publish this data in a sharable form − we could integrate the items of our bank statements into our calendar so that we could see what transactions we made on that day, or include photos so that we could see what we were doing at that time.

However, eight years after publication of this article, we are still some distance realising this vision. In this paper, present an overview of the Semantic Web. We explain why progress has been slow and the reasons we believe this to be about to change.

The paper is organized as follows. In Section II we describe the problems we face when trying to extract meaning from the web as it is today. Section III presents a brief overview of the technologies underlying the Semantic Web. In Section IV we give an overview of the gamut of typical Semantic Web applications and Section V introduces the Linking Open Data project. Finally, we present our conclusions and look to the future in Section VI.

II. THE PROBLEM WITH THE "SYNTACTIC WEB"

In Figure 1 we see a "typical" web page written in HTML which we will use to exemplify some of the drawbacks of the traditional web. This page lists the keynote speeches which took place at the 2009 World Wide Web conference[1]. To the reader, the content of the page can be interpreted intuitively. We can read the titles of the speeches, the names of the speakers and the time and dates at which they take place. Furthermore, by familiarity with browser interaction paradigms, we can realize that by following a hyperlink we can retrieve information about concepts related to the conference (authors, sponsors, attendees etc.). In this example, by following the hyperlink labelled "Sir Tim Berners-Lee" we will retrieve a document containing information about the person of this name. We intuitively assign a meaning - perhaps "has-homepage" - to the hyperlink, allowing us to assimilate the information presented to us.

A web browser cannot assign any to these links we see in this page − a hyperlink is simply a link from one document to another and the interpretation of the meaning of the link (and of the documents themselves!) is a task for the human reader. All that can be inferred automatically is that some undefined association between the two documents exists.

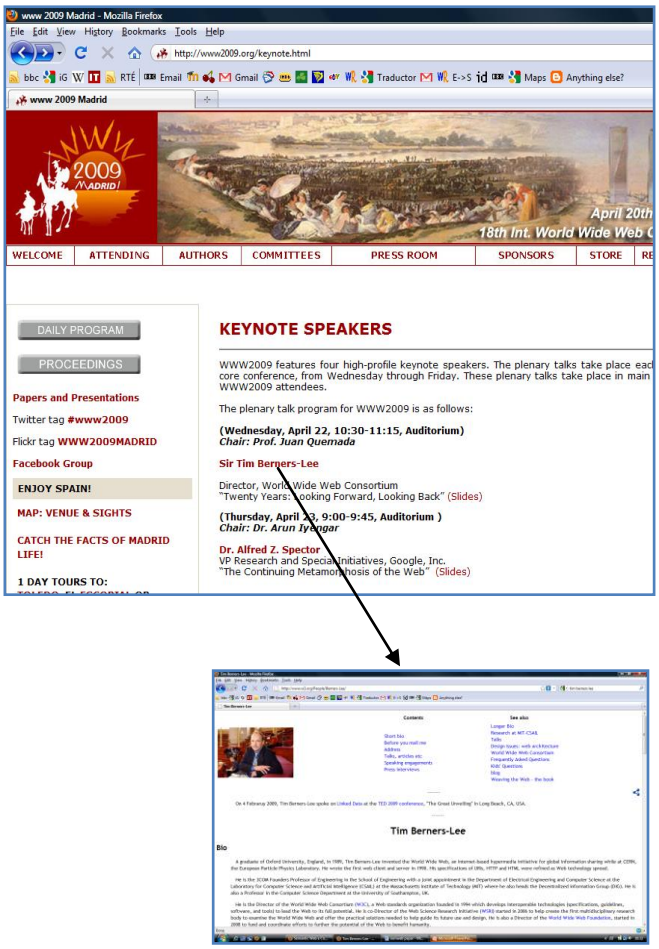

Figure 1. "Traditional" Web Pages with hyperlinks

The problems are even more clear when we consider the nature of keyword-based browsing. While search engines such as Google and Yahoo! are clearly very good at what they do, we frequently are presented with a vast number of results, many (most?) of which will be irrelevant to our search. Semantically similar items will not be retrieved (for instance a search for "movie" will not retrieve results where the word "film" was used). And most significantly, the result set is a collection of individual web pages. Our tasks often require access to multiple sites (such as when we book a holiday), and so it is our responsibility to formulate a sequence of queries to retrieve the individual web pages, each one of which performs part of the task at hand.

There are two potential ways to deal with this problem. One approach is to take the web as it is currently implemented, and to use Artificial Intelligence techniques to analyze the content of web pages in order to provide an interpretation of its meaning. This approach however would be prone to error and would require validation. Furthermore, the rate at which the web is growing would render it practically impossible to achieve.

The other approach is to represent the web pages in a form in which we can represent and interpret the data they contain. If there is a common representation to express the

[1] http://www2009.org/keynote.html

meaning of the data on the web, we can then develop languages, reasoners, and applications which can exploit this representation. This is the approach of the Semantic Web.

III. SEMANTIC WEB TECHNOLOGIES

The Semantic Web describes a web of data rather than documents. And just as we need common formats and standards to be able to retrieve documents from computers all over the world, we need common formats for the representation and integration of data. We also need languages that allow us to describe how this data relates to real world objects and to reason about the data. The famous "Layer Cake" [10] diagram, shown in Figure 2, gives an overview of the hierarchy of the principal languages and technologies, each one exploiting the features of the levels beneath it. It also reinforces the fact that the Semantic Web is not separate from the existing web, but is in fact an extension of its capabilities.

In this section, we summarize and discuss the key aspects shown in the Layer Cake diagram. Firstly we describe the core technologies: the languages RDF and RDFS. Next we describe the higher level concepts, focusing in particular on the concept of the ontology which is at the heart of the Semantic Web infrastructure. Finally we examine the trends and directions of the technology. For further information on the concepts presented in this section, the reader is referred to a more detailed work (eg. [4], [5]).

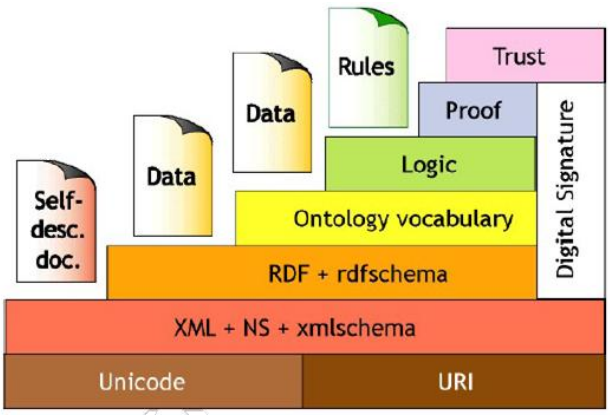

Figure 2. The Semantic Web Layers

A. The Core Technologies: RDF and RDFS

What HTML is to documents, RDF (Resource Description Framework) is to data. It is a W3C standard[2] based on XML which allows us to make statements about objects. It is a data model rather than a language - we can say that an object possesses a particular property, or that it has a named relationship with another object. RDF statements are written as triples: a subject, predicate and object.

By way of example, the statement

"The Adventures of Tom Sawyer" was written by Mark Twain

could be expressed in RDF by a statement such as

```
<rdf:Description
      rdf:about=www.famouswriters.org/twain/mark>
                <s:hasName>Mark Twain</s:hasName>
                <s:hasWritten rdf:resource=
                        www.books.org/ISBN0001047>
</rdf:Description>
```

At first glance it may appear that this information could be equally well represented using XML. However XML makes no commitment on which words should be used to describe a given set of concepts. In the above example we have a property entitled "*hasWritten*", but this could equally have been "*IsAuthorOf*" or another such variant. So, XML is suitable for closed and stable domains, rather than for sharable web resources.

The statements we make in RDF are unambiguous and have a uniform structure. Concepts are each identified by a Universal Resource Identifer (URI) which allows us to make statements about the same concept in different applications. This provides the basis for semantic interoperability, allowing us to distinguish between ambiguous terms (for instance an address could be a geographical location, or a speech) and to define a place on the web at which we can find the definition of the concept.

To describe and make general statements collectively about groups of objects (or classes), and to assign properties to members of these groups we use RDF Schema, or RDFS[3]. RDFS provides a basic object model, and enables us to describe resources in terms of classes, properties, and values. Whereas in RDF we spoke about specific objects such as "“The Adventures of Tom Sawyer” and "Mark Twain", in RDFS we can make general statements such as

"A book was written by an author"

This could be expressed in RDFS as

```
<rdf:Property rdf:ID=“HasWritten”
       <rdfs:domain rdf:resource=“#author”\>
       <rdfs:range rdf:resource=“#book”\>
<\rdf:Property>
```

An expansion of these examples, and the relationship between the graphical representations of RDF and RDFS is shown in Figure 3.

[2] www.w3.org/RDF/

[3] http://www.w3.org/TR/rdf-schema/

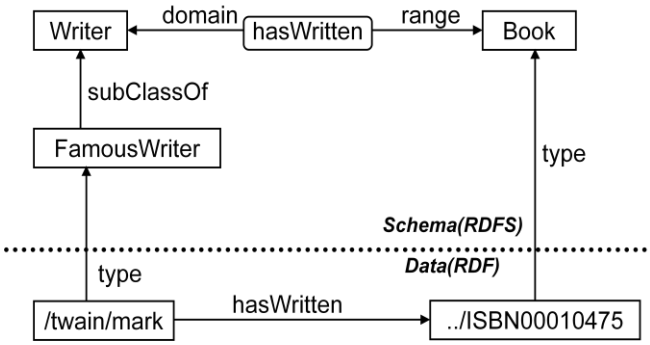

Figure 3. Relationship between RDF and RDFS [5]

B. Ontologies and Reasoning

RDF and RDFS allow us to describe aspects of a domain, but the modelling primitives are too restrictive to be of general use. We need to be able to describe the taxonomic structure of the domain, to be able to model restrictions or constraints of the domain, and to be able to state and reason over a set of inference rules associated with the domain. We need to be able to describe an *ontology* of our domain.

The term ontology originated in the sphere of philosophy, where it signified the nature and the organisation of reality, ie. concerning the kinds of things that exist, and how to describe them. Our definition within Computer Science is more specific, and the most commonly cited definition has been provided to us by Tom Gruber in [6], where he defines an ontology as "an explicit and formal specification of a conceptualization". In other words, an ontology provides us with a shared understanding of a domain of interest. The fact that the specification is formal means that computers can perform reasoning about it. This in turn will improve the accuracy of searches, since a search engine can retrieve data regarding a precise concept, rather than a large collection of web pages based on keyword matching.

In relation to the Semantic Web, for us to share, reuse and reason about data we must provide a precise definition of the ontology, and represent it in a form that makes it amenable to machine processing. An ontology language should ideally extend existing standards such as XML and RDF/S, be of "adequate" expressive power, and provide efficient automated reasoning support. The most widely used ontology language is the "Web Ontology Language", which curiously has the acronym "OWL"[4]. Along with RDF/S, OWL is a W3C standard and augments RDFS with additional constraints such as localised domain and range constraints, cardinality and existence constraints, and transitive, inverse, and symmetric properties.

Adding a reasoning capability to an ontology language is tricky since there will be a trade-off between efficiency and expressiveness. Ultimately it depends on the nature and requirements of the end application, and it is for this reason that OWL offers three sublanguages,

- OWL Lite supports only a limited subset of OWL constructs and is computationally efficient,
- OWL DL is based on a first order logic called Description Logic,
- OWL Full offers the full compatibility with RDFS but at the price of computational tractability.

Examples of applications which could require very different levels of reasoning capabilities are described in the following section.

The top layers of the layer cake have received surprising little attention considering that they are crucial to successful deployment of Semantic Web applications. The proof layer involves the actual deductive process, representation of proofs, and proof validation. It allows applications to inquire why a particular conclusion has been reached, ie. they can give proof of their conclusions. The trust layer provides authentication of identity and evidence of the trustworthiness of data and services. It is supported through the use of digital signatures, recommendations by trusted agents, ratings by certification agencies etc.

C. Recent Trends and Technological Developments

As with any maturing technology, the architecture will not remain static. In 2006 Tim Berners Lee suggested an update to the layer cake diagram [2] which is shown in Figure 4, however this is just one of several proposed refinements. Some of the new features and languages which include the following.

Rules and Inferencing Systems. Alternative approaches to rule specification and inferencing are being developed. RIF (Rules Interchange Format) is a language for representing rules on the Web and for linking different rule-based systems together. The various formalisms are being extended in order to capture causal, probabilistic and temporal knowledge.

Database Support for RDF. As the volume of RDF data increases, it is necessary to provide the means to store, query and reason efficiently over the data. Database support for RDF and OWL is now available from Oracle (although at present the focus is on storage, rather than inferencing capabilities). Other open source products include 3Store[5] and Jena[6]. The specification of a query language for RDF, SPARQL, was adopted by the W3C in 2008.

RDF Extraction. The language GRDDL: ("Gleaning Resource Descriptions from Dialects of Languages") identifies when an XML document contains data compatible with RDF and provides transformations which can extract the data. Considering the volume of XML data available on the web, a means of converting this to RDF is clearly highly desirable.

[4] www.w3.org/2004/OWL

[5] http://sourceforge.net/projects/threestore/

[6] http://jena.sourceforge.net/

Ontology Language Developments. The OWL language was adapted as a standard in 2004. In 2007, work began on the definition of a new version, OWL 2 which includes easier query capabilities and efficient reasoning algorithms scaled to large datasets.

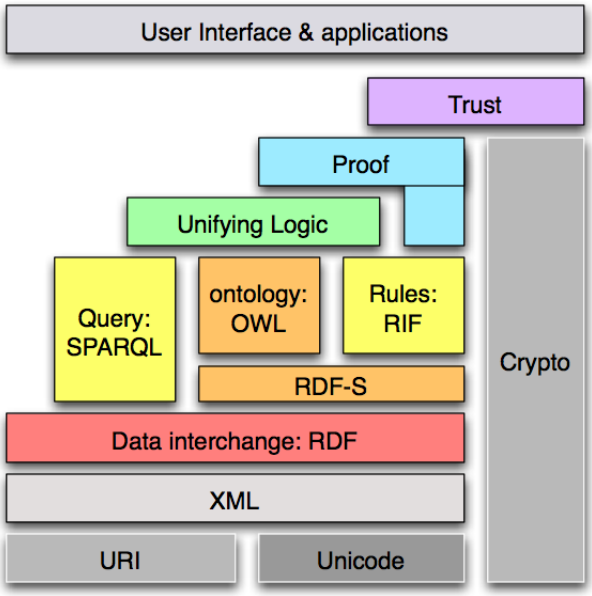

Figure 4. A Revised Semantic Web Layer Cake

IV. The Spectrum of Applications

Even though Semantic Web technology is in its infancy, there are a wide range of applications in existence. In this section we give a brief overview of some typical application areas.

E-Science Applications. Typically e-science describes scenarios involving large data collections requiring computationally intensive processing, and where the participants are distributed across the world. An infrastructure whereby scientists from different disciplines are able to share their insights and results is seen as critical, particularly when we consider the availability of large volumes of data becoming available online. The Gene Ontology[7] is a project aimed at standardizing the representation of genes across databases and species. Perhaps the most famous e-science project is the Human Genome Project[8] which identified the genes in human DNA and which includes over 500 datasets and tools. The International Virtual Observatory Alliance[9] makes available astronomical data from a number of digital archives.

Interoperation of Digital Libraries. Institutions such as libraries, universities, and museums have vast inventories of materials which are increasingly becoming available online. These systems are implemented using a range of different technologies, and although their aims are similar it is a huge challenge to enable the different institutions to access each other's catalogues. Ontologies are useful for providing shared descriptions of the objects, and ontology mapping techniques are being applied to achieve semantic interoperability [3].

Travel Information Systems. The goal of building an application which would allow a user to seamlessly book and plan the various elements of a trip (flights, hotel, car hire etc.) is highly desirable. Ontologies again could be used to arrive at a common understanding of terminology. The Open Travel Alliance is building XML based specifications which allow for the interchange of messages between companies. While this is a first step, an agreed ontology would be needed in order to achieve any meaningful interoperation.

Although many potential applications can be identified, there are less deployed at this time than we might expect. One possible reason is the lack of a common understanding of what the Semantic Web can offer, and more particularly what the role of ontology. At one end of the spectrum we find applications which take the "traditional", or AI view of inferencing, in which accuracy is paramount. Such applications arise in combinatorial chemistry for example, in which vast quantities of information on chemicals and their properties are analysed in order to identify useful new drugs. By coding the required drug's properties as assertions will reduce the number of samples which need to be constructed and manually analyzed by orders of magnitude. In cases such as these, the time taken to perform the inferencing is less important, since the trade-off will be a large reduction in the samples to be analyzed.

At the other end of the spectrum, we have "data centric" web applications which require a swift response to the user. Examples of this type of application include social network recommender systems such as Twine[10] which make use of ontologies to recommend their users to other individuals who may be of interest to them. While it is clear that a response must be generated for the user within a few seconds, we can observe too that there can be no logical proof of correctness and soundness of the answers generated in this type of case! Accordingly, the level of inferencing required in this type of application is minimal.

V. The Future: A Web of Data?

While we have stated that the Semantic Web focuses on data in contrast to the document centric view of the traditional web, this is not the complete picture. In order to realize value from putting data on the web, links need to be made in order to create a "web of data". Instead of having a web with pages that link to each other, we can have (with the same infrastructure) a data model with information on each entity distributed over the web.

The Linking Open Data [3] project aims to extend the collections of data being published on the web in RDF

[7] http://www.geneontology.org/index.shtml
[8] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[9] www.ivoa.net
[10] www.twine.com

format and to create links between them. In a sense, this is analogous to traditional navigation between hypertext documents where the links are now the URIs contained in the RDF statements. Search engines could then query, rather than browse this information.

In a recent talk at the TDC 2009 conference[11], Tim Berners Lee gave a powerful motivation example for the project: scientists investigating the drug discovery for Alzheimer's disease needed to know which proteins were involved in signal transduction and were related to pyramidal neurons. Searching on Google returned 223,000 hits, but no document provided the answer as nobody had asked the question before. Posing the same question to the linked data produces 32 hits, each of which is a protein meeting the specified properties.

At the conception of the project in early 2007, there were a reported 200,000 RDF triples published. By May 2009 this had grown to 4.7 billion [dh]. Core datasets include

- DBpedia, a database extracted from Wikipedia containing over 274 million pieces of information. The knowledge base is constructed by analyzing the different types of structured information, such as the "infoboxes", tables, pictures etc.
- The DBLP Bibliography, which contains bibliographic information of academic papers,
- Geonames, which contains RDF descriptions of 6.5 million geographical features.

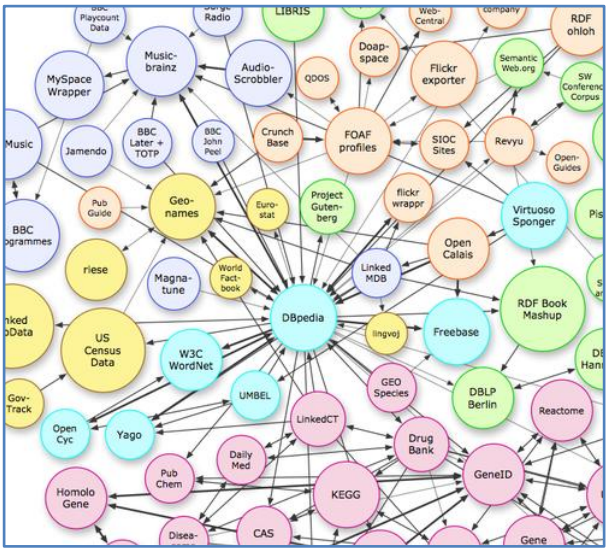

Figure 5. Web of Data (fragment)[12] [July 2009]

VI. Looking Ahead and Conclusions

So where is the Semantic Web? In a 2006 article [11], Tim Berners Lee agreed that the vision he described in the *Scientific American* article has not yet arrived. But perhaps it is arriving by stealth, under the guise of the "Web 3.0" umbrella. Confusion still abounds about the meaning of the term "Web 3.0", which has been variously described as being about the meaning of data, intelligent search, or a "personal assistant". This sounds like what the Semantic Web has to offer, but even if the terms do not become synonymous, it is clear that the Semantic Web will form a crucial component of Web 3.0 (or *vice versa*!).

The last five years have seen Semantic Web applications move from the research labs to the marketplace. While the use of ontologies has been flourishing in niche areas such as e-science for a number of years a recent survey by Hendler [7] shows a marked increase in the number of commercially focused semantic web products. The main industrial players are starting to take the technology more seriously. In August 2008, Microsoft bought Powerset, a semantic search engine, for a reported $100m.

As we have discussed, the "chicken and egg" dilemma is resolving itself with tens of billions of RDF triples now available on the web, and this number is continuing to increase exponentially.

Also, it is becoming easier for companies to enter the market of Semantic Web applications. There are now a wide range of open source applications such as Protégé[13] and Kowari[14] which provide building blocks for application development, making it more cost effective to develop Semantic Web products.

Some observers argue that the Semantic Web has failed to deliver its promise, arguing instead that the Web 2.0 genre of applications signifies the way forward. The Web 2.0 approach has made an enormous impact in recent years, but these applications could be developed and deployed more rapidly as their designers did not have the inconvenience of standards to adhere to. In this article we have demonstrated the steady infiltration from the research lab to the marketplace being made by the Semantic Web over the last decade. As the standards mature and the web of data expands, we are confident that the Semantic Web vision is set to become a reality.

References

[1] Berners-Lee T, Hendler J, Lassila O. 2001. The semantic web. In Scientific American, May 2001, available at http://www.scientificamerican.com/article.cfm?id=the-semantic-web

[2] Berners-Lee et al, 2006. A Framework for Web Science, Foundations and Trends in Web Science. Vol. 1, No 1.

[3] Chen H. 1999. Semantic Research for Digital Libraries, D-Lib Magazine, Vol. 5, No. 10, October 1999. http://www.dlib.org/dlib/october99/chen/10chen.html

[4] Davies J, Fensel D, van Harmelen F (eds). 2003. Towards the Semantic Web: Ontology-Driven Knowledge Management. John Wiley & Sons, Ltd.

[11] http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
[12] http://en.wikipedia.org/wiki/File:Lod-datasets_2009-07-14_colored.png
[13] http://protege.stanford.edu/
[14] http://www.kowari.org/

[5] Fensel D, Hendler JA, Lieberman H, WahlsterW(eds). 2003. Spinning the Semantic Web: Bringing the World Wide Web to its Full Potential. MIT Press: Cambridge, MA. ISBN 0-262-06232-1.

[6] Gruber, T. 1993. Toward principles for the design of ontologies used for knowledge sharing. In Guarino N, Poli R (eds). International Workshop on Formal Ontology, Padova, Italy,

[7] Hendler, J., 2008. Linked Data: The Dark Side of the Semantic Web, (tutorial), 7th International Semantic Web Conference (ISWC08), Karlsruhe, Germany.

[8] Linking Open Data Wiki, available at http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[9] Manning, C., Schütze, H., 1999. Foundations of statistical natural language processing. MIT Press.

[10] "Semantic Web - XML2000, slide 10". W3C. http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html.

[11] Shadbolt, N., Hall, W., Berners-Lee, T., 2006. The Semantic Web Revisited. IEEE Intelligent Systems. http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf.