SEMANTIC WEB

UNFOLDIND THE UNDERLYING TECHNOLOGY

Gaurav Saraf,
F.E., VIIT, Pune.
(+91) 9422500228
gpsaraf@gmail.com

ABSTRACT

The everyday Web has experienced changing trends since it was introduced. We sometimes refer to the current phase of the web as Web 2.0 enriched by community fostered classification and exploitation of information. As technology advances a step further, common and open modeling of data forms the basis of a new web; the "Semantic Web." This simple but radical idea is materialized by importing the principles of knowledge representation from Artificial Intelligence. This paper is an attempt to unfold the technologies that play their indigenous role in the Semantic Layer that would silently position behind the current web as an extension, yet producing unpredictable changes in our interaction with the world. The Semantic Web offers a range of application areas, intelligent automation and real-time scientific publishing being a few. (This futuristic face of the web has been widely referred by the phrase Web 3.0) I also explain here the hurdles that still keep the Semantic Web in research labs, which are however almost at their solution.

KEY WORDS

Semantics, Artificial Intelligence, Web 3.0, Semantic Search, Semantic Publishing, RDF, Web Ontologies.

(I) INTRODUCTION

We may describe the World Wide Web (as we see it today) to be a global set of inter-linked documents. We refer to these documents as WebPages and we may have multimedia components interspersed into them. These documents can be "read" by humans while they are "displayed" by computers. We then find pattern in this data and correlate (or unrelated) it with some other set of data, and this process has its limits being performed almost manually. Today, with HTML and a tool to render it (say, a Web browser or some other user agent), one can create and present a page that lists items for sale. The HTML of this catalog page can make simple, document-level assertions such as "this document's title is 'LookPretty' Superstore". But there is no capability within the HTML itself to assert unambiguously that, for example, item number N802 is a Studio16 Face Wash with a retail price of 70 INR, or that it is a cosmetic product. Rather, HTML can only say that the span of text "N802" is something that should be "positioned" near "Studio16 Face Wash" and "70 INR", etc. There is no way to say "this is a catalog" or even to establish that "Studio16 Face Wash" is a kind of title or that "70 INR" is a price. There is also no way to express that these pieces of information are bound together in describing a discrete item, distinct from other items perhaps listed on the page.

The "Semantic Web" introduces a whole new spectrum of possibilities in this context by adding an "additional layer" of data definitions and relationships behind these documents. The vision of the Semantic Web is to extend the principle of the Web from documents to data.[1] This extension will allow fulfilling more of the Web's potential, in that it will allow data to be shared effectively by wider communities, and to be processed automatically by tools as well as manually. The Semantic Web facilitates deployment of machine power in this correlation and usage of data. At its core, the Semantic Web is comprised of a philosophy, a set of design principles, and a variety of enabling technologies such as the Resource Description Framework (RDF), a variety of data interchange formats, notations and the Web Ontology Language (OWL). [2]

In the next section the underlying philosophy of the Semantic Web is explained. Section (III) elaborates the technologies that make it a reality (RDF, OWL, etc.) This is followed by the applications offered by the Semantic Web and a few problems that it faces.

(II) THE SEMANTIC PHILOSOPHY

The Semantic Web is a "web of inter-related data" (compare this to the phrase, "web of inter-connected documents"). It is an extension to the current World Wide Web in which web content can be expressed not only in natural language, but also in a format that can be "understood" and "used" by automated tools (often called as intelligent agents), thus permitting people and machines to find, share and integrate information more easily. The following occurrences would help us better comprehend the idea: The word "semantic" in a general context would be an adjective for something that makes natural sense, such that decisions can be exercised based upon this sense. In a similar context of computing, "semantic gap" is a phrase used for a distinguishing character between the high-level programming languages and the machine-level language.

The "goals" of the Symantec Web can be summarized as follows:

a) To structure the information over the web as logically inter-related data. (The formatting cues may be placed and rendered separately)
b) To facilitate the use of this sea of data by intelligent software agents collaborating with each other and with their users.
c) To introduce "interoperability" in these relationships also in a way in which they can be used in more than one context.

(III) BUILDING BLOCKS

The Semantic Web is knitted with a set of fundamental building blocks. The following are the technologies developed that help achieve the above stated goals:

a) RESOURCE DESCRIPTION FRAMEWORK (RDF) AND RDF/XML

RDF has evolved as a general method of modeling information, through a variety of syntax formats. It provides a specification to define and describe the relations among data (i.e., resources) on the Web. This is not unlike the usage of hyperlinks on the current Web that connect the current page with another one: the hyperlinks define a relationship between the current page and the target. One major difference is that, on the Semantic Web, such relationships can be established between *any* two resources, there is no notion of "current" page. Another major difference is that the relationship (i.e., the link) itself is *named*, whereas the link used by a human on the (traditional) Web is not and their role is deduced by the human reader. The definition of those relations allow for a better and automatic interchange of data. RDF, which is one of the fundamental building blocks of the Semantic Web, gives a formal definition for that interchange. These resources are usually addressed by a Uniform Resource Identifier (URI) which may or may not begin with *http:* and yet may or may not be accessible via HTTP.

The relationships are expressed in a "subject-predicate-object" manner. The subject of an RDF statement is a resource.

These RDF statements are written in various serialization methods including the XML syntax (denoted by RDF/XML) and the Notaion-3 format.

Suppose we want to assert that the article http://en.wikipedia.org/wiki/India has its "title" as "India" and is published by Wikipedia. In the Notation-3 format, we would write,

```
<http://en.wikipedia.org/wiki/India>
<http://purl.org/dc/elements/1.1/title>
"India"
<http://en.wikipedia.org/wiki/India>
<http://purl.org/dc/elements/1.1/publisher>
"Wikipedia"
```

In this example, http://purl.org/dc/elements/1.1/title is a specific definition for the concept of a title established by the Dublin Core Metadata Initiative which is an example of controlled vocabularies imparted to RDF. This simple data can be utilized (along with many other sets of data) for performing a "semantic search" by intelligent agents in an attempt to say, organize all information that Wikipedia has on India!

Such relationships lead to the formation of a "pseudo-graph" inter-connecting all concerned data. This forms the idea behind Web Ontologies.

b) XML SCHEMA, RDF SCHEMA AND THE WEB ONTOLOGY LANGUAGE (OWL)

The above stated data modeling has resources and relationship amongst these resources that need to be defined and restricted. This is conveniently achieved by certain "Schemas." XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents. RDF Schema is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.

However these vocabularies need to be provided with extensibility and interoperability. OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

c) SPARQL

The SPARQL is a protocol and query language for semantic web data sources. Its name is a recursive acronym that stands for "SPARQL Protocol and RDF Query Language." Compare this to the Structured Query Language (SQL) that is popularly used in the traditional web.

d) THE SERVER SIDE SEMANTICS

The Servers of the Semantic Web would be servers which expose existing data systems using the RDF and SPARQL standards. Many converters to RDF exist from different applications. Relational databases are an important source. The semantic web server components (if addressed in a loose language!) attach to the existing system without affecting its operation.

e) THE CLIENT SIDE SEMANTICS

To experience the Semantic Web, the user would require nothing more than any of today's browsers. This is because the concept of semantic web brings about a complete change in the way information is presented, most of which happens behind the scenes, not necessarily affecting the way information is displayed. It however ushers a new (and in fact unpredictable) user experience.

However the Semantic Web does facilitate the deployment of new "intelligent software agents" that can perform advanced and complex automated tasks that can hardly be imagined in the world of the traditional web.

Also a range of web-based services (often with agents of their own) to supply information specifically to agents may come into the arena (For example, a Trust service that an agent could ask if some online store has a history of poor service or spamming).

(IV) APPLICATION AREAS

The Semantic Web offers many real-world applications that arise from the direct impact it has on the World Wide Web. Real Time "Scientific Publishing" is being considered to be the most important area to benefit from "Semantic Publishing". This directly transforms collaboration trends in Life Sciences and Health Care. Besides other areas expected to go through the wave are:

- Ambient Intelligence
- Semantic Search and Indexing
- Cognitive Systems
- Data Integration
- Multimedia Data Management
- Software Engineering
- Service-Oriented Computing
- Machine Learning
- eScience
- Information Extraction
- Grid Computing
- Peer-to-Peer Systems
- eCommerce
- eGovernment
- Bioinformatics
- Digital Libraries

These again are only imaginable at this point of time as the Semantic Web waits to realize its complete potential.

(V) THE CONSTRAINS

The obvious reason that keeps the Semantic Web still in its evolutionary phase is the challenge that it faces on practical grounds. [3] These hurdles can be summarized in two folds -

A) THE TOOLS PROBLEM

Most of the tools for building semantically-aware applications, or for adding semantics to information are still in the research phase and were designed for expert computer scientists who specialize in knowledge representation, artificial intelligence, and machine learning.

B) THE ONTOLOGY PROBLEM

There are still few widely used or standardized ontologies. And getting people to agree on common ontologies is not generally easy. Furthermore, the world is very complex and to adequately describe all the knowledge that comprises what is thought of as "common sense" would require a very large ontology.

Efforts are at their full momentum at the W3C and many research institutes to make the specifications behind Semantic Web easier to use and smoother to apply. These constrains are almost at the edge of their solution.

The Semantic Web is passing through the same phase through which the early web went. When the "hypertext" was introduced it was a simple yet radical idea that changed the way we lived. The idea behind Semantic Web is equally simple, and perhaps, more radical.

(VI) CONCLUSION

Human Reasoning is a result of certain relationships amongst objects, processes etc. that we have in our minds. The Semantic Web can be said to perform similar reasoning based upon the relationships that we define in web data through RDF. These relationships that we have are also more or less common to many others in the society, which makes our reasoning of practical sense to others.

Similarly a "common vocabulary of relationships" is established for interoperable use through Web Ontologies. The OWL provides a way to enhance this vocabulary. With these and many other basic principles, the Semantic Web, the web of data (in its rightful sense), ushers amazing new ways in which we would use the web.

(VII) REFERANCES

[1] Tim Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, May 2001, pp. 34–43 http://www.sciam.com/article.cfm?id=the-semantic-web

[2] "The Semantic Web" by Wikipedia, at http://en.wikipedia.org/wiki/semantic_web

[3] Ivan Herman (2007). "State of the Semantic Web - Semantic Days 2007"

[4] "The Semantic Web Revisited" by Nigel Shadbolt, Tim Berners-Lee and Wendy Hall, IEEE Intelligent Systems 21(3) pp. 96-101, May/June 2006 - http://eprints.ecs.soton.ac.uk/12614/01/Semantic_Web_Revisted.pdf

[5] FAQs- http://www.w3.org/2001/sw/SW-FAQ

Figure - Understanding the Semantic Web Components (Gaurav Saraf)

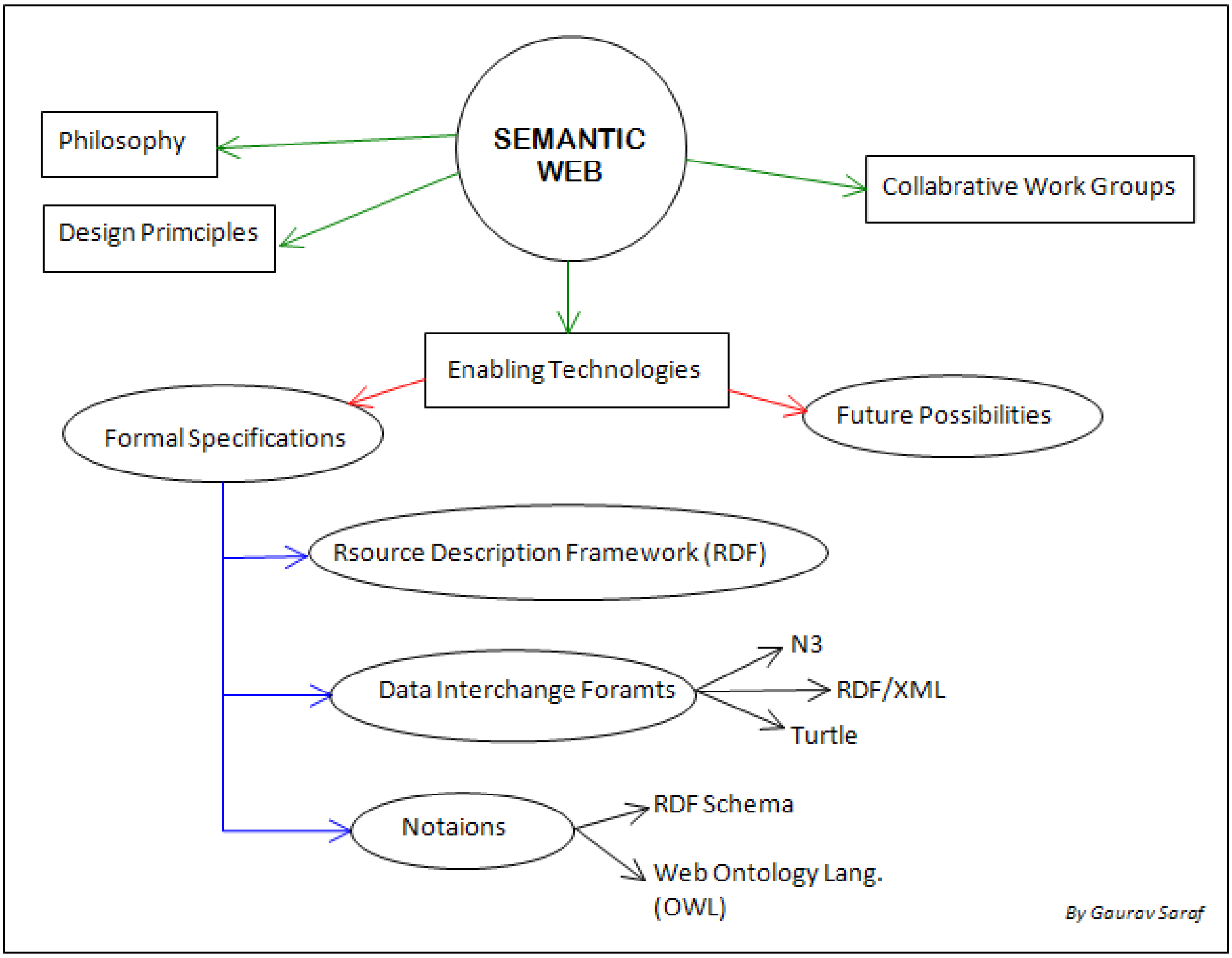

Figure - Emergence of the Semantic Search (by Radar Networks)

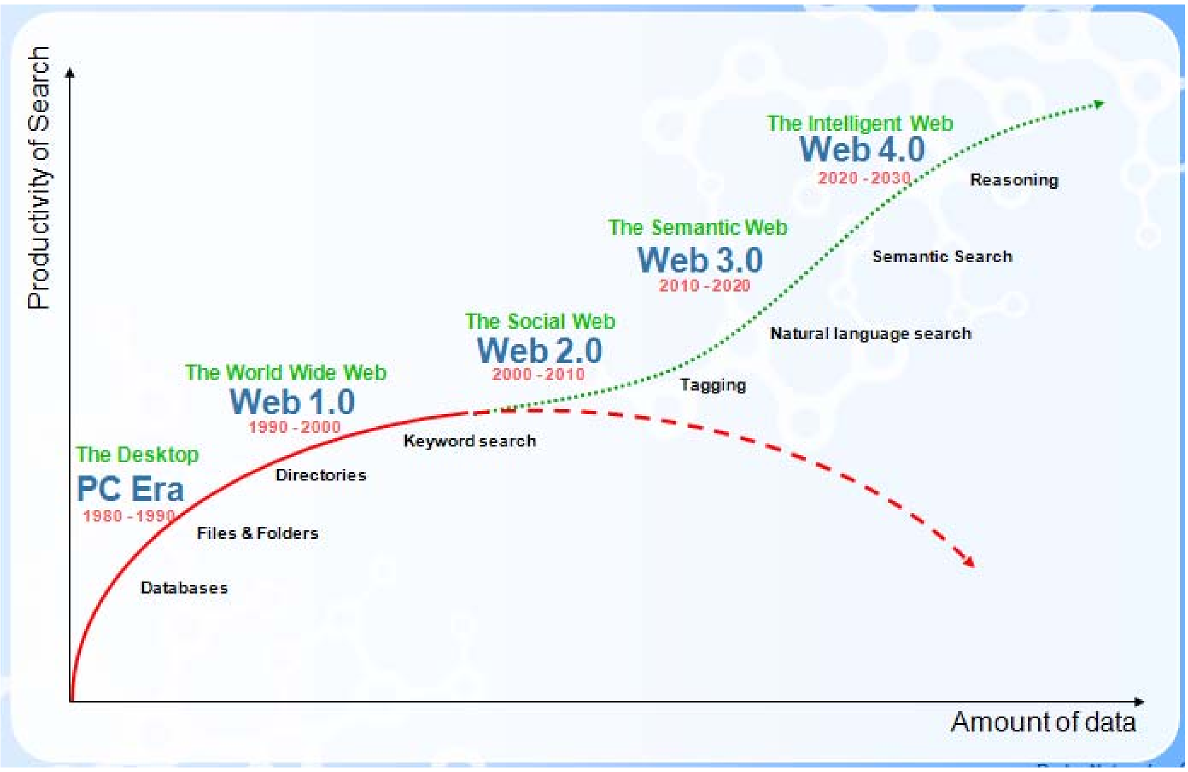

Figure - Generations of the Web (by Radar Networks)