State of the Art of Semantic Web

Ali Alqazzaz
Computer Science and Engineering Department
Oakland University
Rochester Hills, MI 48307, USA
gazzaz86@gmail.com

Abstract

Semantic web is an attempt to provide technology, standards, and methodologies to structure and describe the meaning of data on the web, and to make it easier to process by machines. Thus, allowing software to accomplish many of the tasks users must currently perform manually. The Semantic Web has attracted a diverse, but more important, community of researchers, institutes and companies, all sharing the belief that one day the Semantic Web will have a huge impact on our lives as the current web has. Therefore, there are a lot of work have been done in this area. This paper, gives an overview of Semantic Web and what have been done so far in the Semantic Web filed. Then, it highlights the current major challenges in this field.

Keywords
Semantic Web, Ontology, Data Linked.

I. INTRODUCTION

The World Wide Web was invented by Tim Berners-Lee in 1989. The key technology of the original web—from an end user's point of view, anyway—was the hyperlink. A user could click on a link and immediately go to the document identified in that link. In short, the great advantage of Web 1.0 was that it abstracted away the physical storage and networking layers involved in information exchange between two machines. This breakthrough enabled documents to appear to be directly connected to one another. Click a link and you are there—even if that link goes to a different document on a different machine on another network on another continent. In the same way that Web 1.0 abstracted away the network and physical layers, the Semantic Web abstracts away the document and application layers involved in the exchange of information. The Semantic Web connects facts, so that rather than linking to a specific document or application, you can instead refer to a specific piece of information contained in that document or application. If that information is ever updated, you can automatically take advantage of the update [1,19].

By the definition of World Wide Web Consortium (W3C), “the Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries”[20]. The vision of the Semantic Web is to extend concepts and principles of the Web from documents to data. Therefore, data should be accessed using the general Web architecture using, e.g., URIs; data should be related to one another just as documents are already. Also, this means creation of a common framework that allows data to be shared and reused across application and enterprise to be processed automatically by tools as well as manually, including revealing possible new relationships among pieces of data [1,4,6].

Now, the question is why it is important? Semantic Web technologies help us structure more data and integrate it at Web scale. Thus, it will help avoid reinventing the data description wheel. Also, it helps us describe different types of the information puzzle. As a result, machines can help us to describe ourselves online. Then they can help us to describe places and things in linked ways we have not been able to before. That will help provide contextual search, a form of relational navigation, and ways to infer C from understanding how A and B are connected [1,2,3].

The paper is organized as follows: section II briefly explains the architecture of the Semantic Web. The Semantic Web Technologies are discussed in section III. This is followed by the discussion of the challenges in Semantic Web in section IV. Finally, the paper concludes in section V.

II. SEMANTIC WEB ARCHITECTURE

The basic architecture of the Semantic Web is illustrated in Figure 1 below. Also, it is known as Semantic Web Stack or Semantic Web Layer Cake [21].

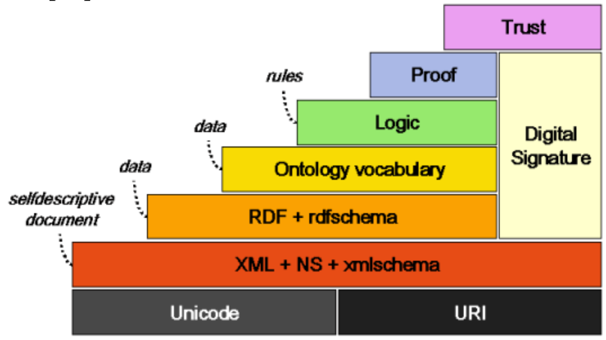

Figure 1. The Semantic Web Stack

- XML: Extensible Markup Language. The language framework that, since 1998, has been used to define nearly all new languages that are used to interchange data over the Web.
- XML Schema: A language used to define the structure of specific XML languages.
- RDF: Resource Description Framework. A flexible language capable of describing all sorts of information and metadata.
- RDF Schema: A framework that provides a means to specify basic vocabularies for specific RDF application languages to use.
- Ontology: Languages used to define vocabularies and establish the usage of words and terms in the context of a specific vocabulary. RDF Schema is a framework for constructing ontologies and is used by many more advanced ontology frameworks.
- Logic and Proof: Logical reasoning is used to establish the consistency and correctness of data sets and to infer conclusions that are not explicitly stated but are required by or consistent with a known set of data. Proofs trace or explain the steps of logical reasoning.

Trust: A means of providing authentication of identity and evidence of the trustworthiness of data, services, and agents.

III. SEMANTIC WEB TECHNOLOGIES

As the previous sections has shown, the term of Semantic Web means the web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. We will classify these technologies based on what they are used for.

A. Linked Data

In order to make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by the SW tools. Furthermore, not only does the Semantic Web need access to data, but relationships among data should be made available to create a Web of Data. This collection of interrelated datasets on the Web can also be referred to as Linked Data [4, 17].

An example of a large Linked Dataset is DBPedia, which, essentially, makes the content of Wikipedia available in RDF. Another example would be FOFA, which is a dataset describing persons, their properties and relationships.

Linked Data builds upon standard Web technologies such as HTTP, RDF and URIs. And, structured data using controlled vocabulary terms and dataset definitions is expressed in Resource Description Framework (RDF) serialization formats such as RDFa, RDF/XML, N3, Turtle, or JSON-LD [4].

The Resource Description Framework (RDF) allows anyone to describe resources, in particular Web resources, such as the author, creation date, subject, and copyright of an image. Both, RDF and its schema language RDFS are recommended by W3C for interlinking resources on the Web and for fostering interoperability among distributed data sources. Therefore, RDF relies on URIs for identifying resources, and constitutes a graph-based data model for linking such resources. To this end, RDF provides the fundamental building blocks for the graph-based data structures that are leveraged by the Semantic Web. More recently, RDF is no longer only the base layer of the Semantic Web, but its importance has increased, and nowadays RDF provides the principal data model for almost all data-minded protocols and formats that are promoted and standardized by W3C[4,6].

RDF graphs are constructed out of RDF triples. . RDF triples describe and connect objects via the combination of resources, properties, and property values. So, the basic data structure of RDF is a triple of the form <subject, predicate, object>. Also, RDF can be represented in different ways such as Triples, XML Trees, N-Triples, and Turtle.

B. Vocabularies

Vocabularies in the Semantic Web define the concepts and relationships used to describe and represent a certain domain. They are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only). So, they are the basic building blocks for inference techniques on the Semantic Web [6, 17].

The main roles of vocabularies in the Semantic Web are to help data integration when ambiguities may exist on the terms used in the different data sets, or when a bit of extra knowledge may lead to the discovery of new relationships. Therefore, W3C offers a large palette of techniques to describe and define different forms of vocabularies in a standard format. These include RDF and RDF Schemas, Simple Knowledge Organization System (SKOS), Web Ontology Language (OWL), and the Rule Interchange Format (RIF) [17].

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies [17]. The OWL languages are characterized by formal semantics. They are built upon a W3C XML standard for objects called the Resource Description Framework (RDF). On the other hand, SKOS is a common data model for sharing and linking knowledge organization systems via the Web. The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive language for developing and sharing new knowledge organization systems. It may be used on its own, or in combination with formal knowledge representation languages such as the Web Ontology language (OWL) [6, 19].

C. Query

Query in the Semantic Web context means technologies and protocols that can programmatically retrieve information from the Web of Data. RDF is the data model for Semantic Web data, and SPARQL is the standard query language for this data model.

SPARQL is essentially a graph-matching query language. A SPARQL query is composed of: (1) a body, which is a complex RDF graph pattern-matching expression, and (2) a head, which is an expression that indicates how to construct the answer to the query [6, 10, 17]. Technically, SPARQL queries are based on (triple) patterns. RDF can be seen as a set of relationships among resources (i.e., RDF triples); SPARQL queries provide one or more patterns against such relationships. These triple patterns are similar to RDF triples, except that one or more of the constituent resource references are variables. A SPARQL engine would returns the resources for all triples that match these patterns [17].

D. Inference

Generally, inference on the Semantic Web can be characterized by discovering new relationships. On the Semantic Web, data is modeled as a set of (named) relationships between resources. “Inference” means that automatic procedures can generate new relationships based on the data and based on some additional information in the form of a vocabulary, e.g., a set of rules.

In the Semantic Web, the source of such extra information can be defined via vocabularies or rule sets. Both of these approaches draw upon knowledge representation techniques. In general, ontologies concentrate on classification methods, putting an emphasis on defining 'classes', 'subclasses', on how individual resources can be associated to such classes, and characterizing the relationships among classes and their instances. Rules, on the other hand, concentrate on defining a general mechanism on discovering and generating new relationships based on existing ones, much like logic programs do.

The core idea of inference on the Semantic Web is to improve the quality of data integration on the Web, by discovering new relationships, automatically analyzing the content of the data, or managing knowledge on the Web in general. Inference based techniques are also important in discovering possible inconsistencies in the (integrated) data [6, 17]. Rule Interchange Format (RIF) has been developed to cover rule based approaches. The goal of RIF is to define a standard for exchanging rules among rule systems especially among Web rule engines. RIF focuses on exchange rather than defining a single one-fits-all rule language [22].

IV. CHALLENGES ON SEMANTIC WEB

In spite of the fact that a lot of work have been done in the SW field, there are some serious problems. In the following, we will highlight the major issues to the best of our knowledge.

A. Availability of Content

In order to take advantage of the concept of Semantic Web, the existing web content should be upgraded to Semantic Web content. That includes static HTML pages, existing XML content, and dynamic content, multimedia and web services. Since the infrastructure of the Semantic Web is still being built (RDFS, OIL, DAML+OIL, etc.), there is little Semantic Web content available. Apart from the infrastructure, researchers are currently building tools to support semantic annotation of web content. Such tools are important and critical to the success of the Semantic Web [11].

B. Ontology

This challenge is related to ontologies availability, development and evolution. As we have seen, ontologies play main role in the SW because they are the carriers of the meaning contained in the SW. An effort must be done in having common widely used ontologies for the Semantic Web, on the provision of adequate infrastructure for ontology development, change management and mapping, and, in this distributed web environment, on the adequate control of the evolution of ontologies and the annotations referring to them [11].
One of the issues is the evolution of ontologies and their relation to already annotated data. Configuration management tools are necessary to keep control of the versions of each ontology as well as the interdependencies between them and annotations. The second issue is the construction of basic or kernel ontologies to be used by all the domains.

C. Multilinguality

This issue already exists in the traditional Web, and should also be tackled in the Semantic Web. However, the growth of semantic information has been stimulated to a large extent by the emergence of Linked Data. Although this brings us a step closer to the vision of the Semantic Web, it also raises new issues such as the need for dealing with information expressed in different natural languages. Indeed, although the Web of Data can contain any kind of information in any language, it still lacks explicit mechanisms to automatically reconcile such information when it is expressed in different languages. This leads to situations in which data expressed in a certain language is not easily accessible to speakers of other languages [14].
As a result, any Semantic Web approach should provide facilities to access information in several languages, allowing the creation and access to SW content independently of the native language of content providers and users.

D. Scalability of Semantic Web Content

After we have the SW content, we need to worry about how to manage it in a scalable manner, how to organize it, where to store it and how to find the right content. So, all these tasks must be performed and coordinated in a scalable manner, as these solutions should be prepared for the huge growth of the Semantic Web [11]. For example, one issue is related to the storage and organization of Semantic Web pages. The 'basic' Semantic Web consists of ontology-based annotated pages whose linking structure reflects the structure of the WWW, that is, pages connected to others by means of hyperlinks. This hyperlinked configuration does not fully exploit the underlying semantics of Semantic Web pages. We foresee the use of semantic indexes to group Semantic Web content based on particular topics. This is a necessary step to make applications able to aggregate content in order to provide added value services. Semantic indexes will be generated dynamically using ontological information and annotated documents [11].

E. Trust, Security, and Privacy

Semantic Web applications assume and expect that the information content of resources is of high quality and can be trusted. Similarly, security and privacy of sensitive information on the SW must be insured. A lot of work must be made in developing comprehensive solutions to ensure the trust, security, and privacy of Semantic Web content.

V. CONCLUSION

The Semantic Web is an initiative that aims at improving the current state of the World Wide Web. And, the key idea is the use of machine-processable Web information. Semantic Web technologies as a whole have made tremendous strides in the last decade. Key technologies include explicit metadata, ontologies, logic and inference, and intelligent agents. Thus, the development of the Semantic Web proceeds in layers. The Semantic Web standards such as RDF, SPARQL, OWL, and others were merely drafts in 2001, but they have now been formalized and ratified.

In spite of significant early corporate adoption by a select few frontrunners, most companies have not yet started using (or are even unaware of the existence of) Semantic Web technologies. The learning curve for using Semantic Web technologies is steep because few educational resources currently exist for users new to the concepts, and still fewer resources can be found that discuss when and how to apply the technologies to real world scenarios.

References

[1] T. Berners-Lee, J. Hendler and O. Lassila. The Semantic Web. Scientific American, May 2001.

[2] JQ Anderson, H Rainie. The fate of the Semantic Web, Pew Research Center. (2010)

[3] Cardoso, J. The Semantic Web vision: where are we? IEEE Intelligent Systems, vol. 22, no. 5, September/October 2007, pp. 84-88

[4] J. Domingue; D. Fensel; J. Hendler. Handbook of semantic web technologies, Springer, 2011, p. 1076

[5] J. Hendler, J. Golbeck. Metcalfe's law, Web 2.0, and the Semantic Web. Elsevier B.V, 2008, Volume 6, Issue 1, pp. 14 – 20

[6] G. Antoniou, F. Harmelen. A Semantic Web primer, MIT Press: Cambridge, MA, 2008.

[7] T. Berners-Lee. Long Live the Web: A Call for Continued Open Standards and Neutrality, Scientific American, 22 November 2010

[8] GL. dos Santos. "THE SEMANTIC WEB." (2011)

[9] B. Middleton, J. Halbert, and F. Coyle. Security Impacts on Semantic Technologies in the Coming Decade.

[10] V. Sugumaran, and J. A. Gulla. Applied semantic web technologies. CRC Press, 2011.

[11] J. Contreras, C. Oscar, and A. Gómez-Pérez. Six Challenges for the Semantic Web. 2009.

[12] N. Shadbolt, H. Wendy, and T. Berners-Lee. The semantic web revisited. IEEE Intelligent Systems, v.21 n.3, p.96-101, 2006.

[13] A Gómez-Pérez, O Corcho. Ontology languages for the semantic web. IEEE Intelligent Systems, v.17 n.1, p. 54-60, 2002.

[14] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez a, P. Buitelaar c, J. McCrae. Challenges for the multilingual web of data. Web Semantics: Science, Services and Agents on the World Wide Web. v.11, p.63-71, 2012.

[15] M. Arenas, J. Pérez. Querying semantic web data with SPARQL. Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2011.

[16] M. Farouk, M. Ishizuka. Semantic Conversion for Dynamic Web Pages. International Conferences on Web Intelligence and Intelligent Agent Technology. v.3, p.285-288. 2011.

[17] W3C Semantic Web Activity. W3C. Retrieved 24 March 2015 from: http://www.w3.org/standards/semanticweb/

[18] M. Dabrowski. Semantic Web. The International Encyclopedia of Digital Communication and Society.v.1, p.1–4, 2015.

[19] Cambridge Semantics. Semantic Web [Online]. Available: http://www.cambridgesemantics.com/semantic-university/introduction-semantic-web Retrieved in 03/21/2015.

[20] W3C. Semantic Web[Online]. Available: http://www.w3.org/2001/sw/ Retrieved in 03/21/2015.

[21] A Harth, M Janik, S Staab. "Semantic web architecture." Handbook of Semantic Web Technologies. Springer Berlin Heidelberg, 2011. 43-75.

[22] P Hitzler, M Krotzsch, S Rudolph. Foundations of semantic web technologies. CRC Press, 2011.

Biography

Ali Alqazzaz is currently a student at Oakland University, Michigan, USA. He earned B.S. in Computer Science from King Khalid University, Saudi Arabia, Masters in Computer Science from Saint Joseph's University, USA. He has published conference papers. His research interests include semantic web, information security, optimization, and parallel computing. He is student member of IEEE.