Creating Knowledge and Semantic Web Applications on the Web

M. Anwar Hossain and Abdulmotaleb El Saddik
Multimedia Communications Research Laboratory
University of Ottawa
{anwar, abed}@mcrlab.uottawa.ca

Abstract

The huge quantity of information already available on the current web could be integrated in a more meaningful way. This has raised new challenges to building a semantic web infrastructure where documents will be understandable not only by humans but also by computers, information will have semantic meanings, and software agents will be able to automatically access the underlying knowledge and process it to assist humans. In spite of significant work on semantic data models, languages, ontologies, agents, and search engines, knowledge creation and development of semantic web applications has not yet been easy. This paper emphasizes on using the existing semantic web technologies and tools in order to create knowledge and semantic web applications on the Web.

Keywords: Semantic web, semantic web languages, semantic web tools.

1. Benefits of the Semantic Web

In the semantic web, metadata are used to describe web pages, documents, databases, models, concepts, and other web resources. This gives software applications and agents an understanding of the knowledge within the content, which will provide substantial benefit to the whole web community. Some of these benefits are: improved web search facilities; adaptive user interface; enhanced collaborative filtering of information; convenient web services; flexible data integration from multiple domains; efficient knowledge management; and infer additional facts that are not explicitly made.

2. Adding Semantics: The Challenge

Today, the web has become a very important part of most people's life. Its success is based on its simplicity and availability. However for the phenomenal growth of web population (Fig. 1), it becomes increasingly difficult to organize, locate and integrate knowledge available on the web. As a consequence, the huge volume and simplicity has become an obstacle to its further growth.

In the current Web, computers have dummy access to the actual knowledge contained in text/html pages, documents, music files, images, etc. They offer little support to users in accessing and processing this information. Semantic web development can bring meaningful structure to the content of web pages, and software applications or agents can then carry out sophisticated operations for individual users. Then the web will not only provide pages and links, but also relationships between the actual content. For example, an online semantic web vocabulary could describe that a postal code is equivalent to a zip code when referring to an address. Overall, computers will cease just being a 'dummy' terminal for information input and output. It would be able to carry out automated tasks for users with a high-level description of the task's objective.

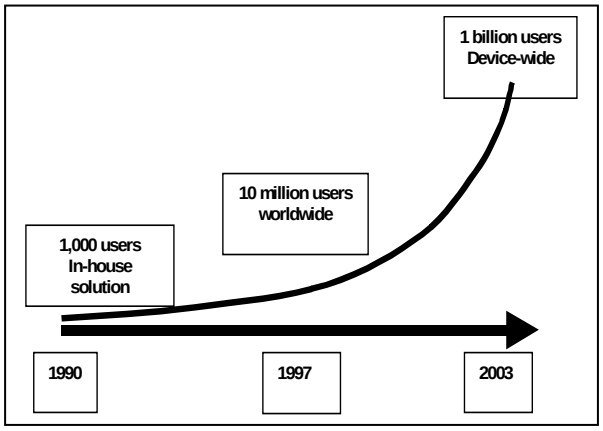

Fig. 1. The growth of the web [13]

The semantic web faces two main challenges towards its growth. The first is the effort to link existing content to semantic meaning by using some sort of metadata. The challenge is to enable the general users who are not expert in logic to create machine-understandable semantic content. This can be addressed by providing an easy way to mark up content with semantic tags, which should be done automatically and should be lower in cost or even free [17]. The second challenge is to develop a set of applications that make use of this newly generated metadata-based knowledge. To attract more population into the development of semantic web, new and interesting semantic web applications should be developed by the already committed participants.

3. Semantic Web Languages

Tim Berners-Lee at the XML 2000 conference has presented a layered architecture (Fig. 2) of the semantic web along with proposed languages at each of these layers. The foundation of the layers is based on Unicode, URI and XML technology. We assume that the reader is familiar enough with Unicode and Uniform Resource Locators (URLs), which are similar to URIs (Universal Resource identifiers) and thus no emphasis will be laid on this basic aspect.

XML is the cornerstone for the second layer and can be used to describe in a convenient way the data stored on the web. However, XML cannot be used directly to express relationships among several pieces of data in a machine-understandable way. Thus, Resource Description Framework (RDF) [16] is introduced. RDF is not a true language, but rather a web metadata model. RDF can be syntactically expressed in XML, known as RDF/XML [9], and offers the possibility of creating semantic content. Some other alternative interchange formats for RDF are N3 [29] and N-triples [10], which are considered a simpler way of understanding the RDF model. However XML as a widely accepted format is likely to dominate.

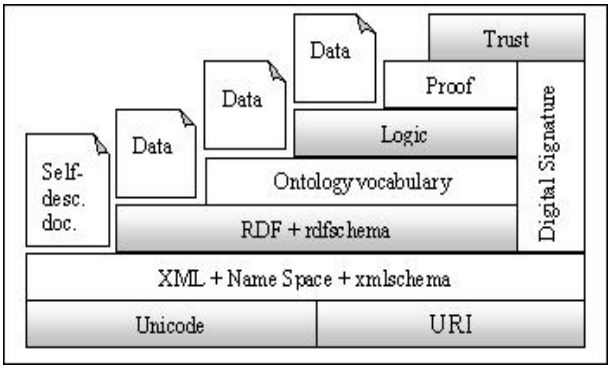

Fig. 2. Semantic web layer cake by Tim Berners-Lee

The RDF data model is too generic and less expressive. It uses URI to identify terms, but does not say what those terms mean. As a consequence, RDF-Schema [8] has evolved on top of RDF to define the meanings and relationships of terms and group them into classes and to assign them appropriate properties.

The next upper layer is the ontology vocabulary layer. Ontology may be defined as "set of knowledge terms, including the vocabulary, the semantic interconnections and some simple rules of inference and logic" [17]. The goal of ontology is to enable wide semantic interoperability, to allow the computers to share the meanings embedded in the metadata, and facilitate machine-to-machine communication using software agents. To gain even more expressive power over RDF-Schema, new ontology languages have emerged. Logically RDF-Schema is a kind of primitive ontology. There are also other ontology languages like SHOE [18], Ontobroker [14], and OIL [15]. The combination of the best features of RDF, SHOE, and OIL has led to the creation of DAML+OIL [12] language, which is also based upon RDF. This DAML+OIL language is the basis of the current recommendation of W3C Ontology Web Language (OWL) [11].

Above the ontology layer there is logic framework, which will enable semantic web-based engines to deduce new knowledge from the existing knowledge. This will allow software agents to bring proof by following the semantic links and monitoring the logic steps [1]. Accordingly a rating of sources and processes will be available online. The question of trusting those ratings will remain open. Digital signature comes into play to ease the situation by identifying the author of the statements and sources. When all these are in place, people will be able to trust the web in terms of the information they see, the facts they perceive, and the claims they receive thereby creating a "web of trust".

4. Building the Semantic Web

Several steps are involved in building semantic web application. The first step is to create semantic web content to gather knowledge from existing and new web documents. This will involve marking up the content with semantic tags using semantic web languages and linking them to online ontology. The creation of ontology for different knowledge domains is a challenge and requires corresponding subject matter and logic experts [17]. The fact emphasizes the use of existing ontology. FOAF (http://rdfweb.org) is a popular vocabulary for RDF/XML, which provides an easy way to describe people and links between them. The DAML web ontology library (http://www.daml.org/ontologies/) lists several other online ontologies that can be linked to create semantic content. The second step is concerned with validating the newly generated content. Once the content is ready, the third step will involve the use of this content by software agents to search, infer facts, and do other interesting tasks in favour of humans. We will highlight these steps in further details in the following sections.

4.1 Creating Semantic Web Contents

Content creation requires new semantic content using a suitable semantic web language, as well as handling existing content already available in abundant on the web. The challenge is to give semantics to the existing content. There are many formats which can be grouped as follows: HTML documents; XML

documents; Text files; Data within database records; Image files/ Digital photos, e.g. JPEG, TIF, WMF; Music files, e.g. MP3; Email messages, e.g. Outlook emails; Other multimedia content. We will describe how each of these diverse sets of data can be marked up semantically using existing semantic web tools and technologies.

HTML/ XHTML documents are described by tags, which are only good for displaying the embedded content in a user-readable fashion. However there are several approaches [26] to put semantics in these documents including embedding or linking RDF metadata in HTML/ XHTML documents. Embedding has been adopted by Mangrove [21], where its Graphical Tagger Tool or any other text editors can be used to insert semantic tags inside HTML documents. The tag semantics are defined in a schema beforehand. All the embedded annotations are directly converted into RDF by using the Jena [6] toolkit to do all the processing in RDF.

Plain and delimited text files containing data can be converted to RDF using the ConvertToRDF [19] tool. This data is normally generated from spreadsheets or databases; mapping the column headings of the data to online ontology produces the semantic RDF data.

The huge collection of XML documents could contribute heavily to the creation of knowledge on the Web. Although there are some syntactic differences between pure XML and the RDF serialized XML (RDF/XML), existing XML documents can be made RDF-friendly by following some guidelines stated in [5]. Some of the guidelines include using specific namespace for Elements, using URI to refer existing ontology instead of creating one, using rdf:ID attributes and so on.

Relational databases consist of huge number of structurally formatted data. D2R MAP [7] is a Database to RDF Export tool. The D2R processor can export data from relational databases in RDF, N3 or N-triples format and uses an XML-based declarative mapping language to map between a database schema and a RDF schema or OWL ontology.

Image files and digital photos can be semantically illustrated using RDFPic (http://jigsaw.w3.org/rdfpic/), a tool that follows the metadata model described in [30], by embedding the RDF description of the image into the image. This description is separated into three other metadata schemas: the RDF format of Dublin Core [25] schema, which describes creator, editor, title, date of publishing, publisher and other metadata about the image; the technical schema, which describes technical information such as type of camera, type of film, date, and scanner model; the content schema, which describes portrait, group portrait, landscape, architecture, sport, animals, and other content related information. One obvious benefit of storing such metadata about an image will be to quickly and easily find a particular image from a larger list.

The semantic indexing application MusicBrainz [2] enables user to store and exchange metadata of media files on the Internet. Music metadata for digital audio (e.g. mp3) and video tracks such as artist's name, album name, tracks and so on are represented using RDF/XML format and shared among the music players on the web. The MusicBrainz tagger has made it easy to find the right music from a huge collection.

Another approach is Annotea [20], which provides an easy way to create annotation about online documents. Annotations such as comments, notes, typographical corrections, hypotheses and other metadata are stored in generic RDF databases on a separate annotation server. This can later be accessed to search the actual document, as well as perform other semantic based tasks.

Table 1 summarizes the above-discussed tools and/or technologies that can be leveraged to add semantics to existing web content.

Table 1. Tools/technologies used to add semantics to existing web contents

Web contents	**Method used to add semantics**	**Tools/technologies used**
HTML/ TEXT	- Embed/ link RDF data	Manual process
	- Embed semantic tag	Mangrove graphical tagger
XML document	- Restructured to make RDF-Friendly	Manual process
Delimited Text File generated from spreadsheet/database	- Mapping column headings to online ontology	ConvertToRDF
Database Record	- Declarative mapping	D2R Map
Image files	- Embed RDF description into the image	RDFPic
Music files	- Stores music metadata in separate server - Inserts music metadata into the music files	MusicBrainz
Online web document in any format	- Adding comments, notes, explanations, or other types of external remarks in separate annotation server	Annotea

To create new semantic web contents there are several tools that can be used at the moment. Some of these tools are Protégé-2000 [22], the RDF Editor in [3], and Jena [6].

Protégé-2000 [22] is a java tool for knowledge model design and acquisition. It supports the creation of RDF schema, RDF data, and OIL ontology. Its visual interface makes it easy for the user to concentrate on the knowledge model and is capable of translating a model from one language to another. Protégé-2000 is currently customized to support editing the ontology written in OWL, a new ontology language proposed by the W3C.

The RDF Editor described in [3] provides an integrated environment for semantic markup of HTML pages by borrowing terms from existing online ontologies. It has a simple graphical user interface that helps users to markup their documents with minimal knowledge of RDF terms and syntaxes. Users can identify words and phrases in the HTML document to be included as RDF triples and save it as RDF/XML file. The ontologies that are used can be stored in a local database for subsequent access.

The Jena toolkit [6] is a Java framework for building semantic web applications. It has a RDF API that supports new RDF data types. Jena supports writing RDF in RDF/XML, N3 and N-Triples and is capable of parsing these formats. It has an ontology API, which basically supports major semantic web languages like OWL, DAML+OIL, and RDF-S. Jena toolkit supports both in-memory and persistent storage (e.g. MySQL, Oracle, and PostgreSQL databases) of RDF data. Its persistent storage subsystem is open to supporting other SQL databases as well.

4.2 Validating the Semantic Contents

The next step after creating semantic contents is to validate RDF/XML data. The valid data will enable search engines and other software agents to find, use and re-use these contents. The most commonly used RDF utility is the RDF Validator [4], a tool to parse RDF/XML contents to ensure that it's valid, as well as to generate different views of the data model including creating an image of the RDF graphs and showing the results as triples.

Another tool is the ICS-FORTH Validating RDF Parser (VRP), which can be used to validate RDF statements in RDF/XML documents [24]. The validation is done against a stored RDF schema according to the current RDF model and specification. VRP also supports XML schema datatypes and Unicode character set and provides a flexible means to activate or deactivate semantic constraints on which the validation would be performed.

4.3 Using the Semantic Contents

The third step is to develop semantic services utilizing the semantic content that is marked with semantic tags and connected to online ontology. Software agents will parse this content and deliver several services to the user. Some of these services include improved semantic search, appointment books, trip organizing, hotel seat reservation, auction negotiating, and meeting scheduling and so on. The ultimate goal of the semantic web would be to realize such a vision as depicted in [17, 28] and others.

As an example, RETSINA Calendar Agent [27] is a distributed meeting scheduler, which can parse the semantically marked-up meeting schedule and determine if there is any conflict between the existing appointment event and the potential new appointment event. RESTINA has the capability to request meetings with one or more individuals for a given time slot; provides interoperability between RDF-based calendar descriptions on the web and Personal Information Manager (PIM) Systems such as Microsoft's Outlook 2000; and is capable of importing selected schedules into MS Outlook and then sending email messages to the attendees.

Perhaps the most important use of semantic web agents is to search the semantic contents and answer user's query. Semantic Search [23] is an application that includes Activity Based Search (ABS) and W3C Semantic Search systems. The ABS search is intended for larger domains of data about musicians, athletes, actors, places, and products already present in different web sites. However, due to the unavailability of machine-understandable data of these domains, an HTML scrapper is used to locate and convert the human readable data from those web sites into machine-readable data. The knowledge base of the application framework also provides data for ABS system. These data are then queried using a simple GetData query interface provided by TAP. Unlike ABS system, the W3C Semantic Search is based on the data of its web site, which includes people, W3C Activities, Working Groups and other Communities, Documents, and News. Part of these data is available in RDF/XML and others can be converted to the same format, and can be searched semantically. The semantic data in both the search systems are linked to respective ontology provided by the framework.

The Jena toolkit [6] provides RDQL query language to query the RDF content. Any search agent can use this query language to provide search facilities for the semantic web. RDQL has a similar syntax to standard SQL, making it attractive to the search agent developers.

The above discussion emphasizes the fact that semantic web agents will open new paradigm of useful

services for information interchange and interaction between human-computer, computer-computer, and business-business. The agents will help us to accomplish our work in a more efficient way.

5. Conclusion

Knowledge creation and the development of semantic web applications are interesting only if it becomes widespread, with ontologies available everywhere for all knowledge domains. The major hurdle now is the lack of support from industry as well as personal users. To address this issue we are developing a specialized semantic web application to demonstrate the potential of the semantic web in a multi-domain environment. This article is the result of our initial investigation of the tools and technologies already available that can be used to create knowledge in terms of semantic web content from existing and new web resources. This would lead the web community a step further towards building the semantic web.

References

[1]. A. Swartz and J. Hendler, "The Semantic Web: A Network of Content for the Digital City," *Proc. 2nd Ann. Workshop on Digital Cities* (Oct. 2001, Kyoto).

[2]. A. Swartz, "MusicBrainz: A Semantic Web Service," *IEEE Intelligent Systems*, vol. 17, no. 1, Jan./Feb. 2002, pp. 76-77.

[3]. A. Kalyanpur et al., "An RDF Editor and Portal for the Semantic Web," *Workshop on Semantic Authoring, Annotation & Knowledge Markup (SAAKM)* at ECAI 2002 (July 2002, Lyon, France).

[4]. A. Barstow, *W3C RDF Validation Service*, W3C/HP. http://www.w3.org/RDF/Validator/.

[5]. B. DuCharme and J. Cowan, *Make your XML RDF-Friendly*, Oct. 30, 2002, http://www.xml.com/pub/a/2002/10/30/rdf-friendly.html?page=1

[6]. B. McBride, "Jena: A Semantic Web Toolkit," *IEEE Internet Computing*, vol. 6, no. 6, Nov./Dec. 2002, pp. 55-59.

[7]. C. Bizer, "D2R MAP - A Database to RDF Mapping Language," (poster), *Proc. 12th Int'l Conf. World Wide Web* (WWW 03, May 2003, Budapest, Hungary), http://www2003.org/cdrom/papers/poster/p004/p4-bizer.html.

[8]. D. Brickley and R.V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Recommendation, Feb. 2004, http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[9]. D. Beckett, *RDF/XML Syntax Specification (Revised)*, W3C Recommendation, Feb. 2004, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

[10]. D. Beckett, *N-Triples W3C RDF Core WG Internal Working Draft*, 2001, http://www.w3.org/2001/sw/RDFCore/ntriples/.

[11]. D.L. McGuinness and F.V. Harmelen, *OWL Web Ontology Language Overview*, Feb. 2004, http://www.w3.org/TR/2004/REC-owl-features-20040210/.

[12]. D.L. McGuinness, J. Hendler, and L.A. Stein, "DAML+OIL: An Ontology Language for the Semantic Web," *IEEE intelligent systems*, vol. 17, no. 5, Sept./Oct. 2002, pp. 72-80.

[13]. D. Fensel and M.A. Musen, "The Semantic Web: A Brain for Humankind," *IEEE Intelligent Systems*, vol. 16, no. 2, Mar./Apr. 2001, pp. 24-25.

[14]. D. Fensel et al., "Ontobroker: How to make the WWW Intelligent," *AAAI-98 Workshop on AI and Information Integration*, AAAI Press, Menlo Park, CA., 1998, pp. 36-42.

[15]. D. Fensel et al., "OIL: An Ontology Infrastructure for the Semantic Web," *IEEE Intelligent Systems*, vol. 16, no. 2, Mar./Apr. 2001, pp. 38-45.

[16]. F. Manola and E. Miller, *RDF Primer*, W3C Recommendation, Feb. 2004, http://www.w3.org/TR/2004/REC-rdf-primer-20040210/.

[17]. J. Hendler, "Agents and the Semantic Web," *IEEE Intelligent Systems*, vol. 16, no. 2, Mar./Apr. 2001, pp. 30-37.

[18]. J. Heflin, J. Hendler, and S. Luke, "SHOE: A Blueprint for the Semantic Web," *Spinning the Semantic Web*, D. Fensel et al., eds., MIT Press, 2003, Cambridge, MA.

[19]. J. Golbeck et al., "New Tools for the Semantic Web," *Proc. 13th Int'l Conf. Knowledge Eng. and Knowledge Management* (EKAW 02, Oct. 2002, Siguenza, Spain), http://www.ece.umd.edu/~adityak/EKAW02.pdf.

[20]. J. Kahan et al., "Annotea: An Open RDF Infrastructure for Shared Web Annotations,"

Proc. 10th Int'l Conf. World Wide Web (WWW 01, May 2001, Hong Kong).

[21]. L. McDowell et al., *Evolving the Semantic Web with Mangrove*, tech. report TR-03-02-01, Computer Science and Eng. Dept., Univ. of Washington, Seattle, WA., 2003.

[22]. N.F. Noy et al., "Creating Semantic Web Contents with Protégé-2000," *IEEE Intelligent Systems*, vol. 48, no. 2, Mar./Apr. 2001, pp. 60-71.

[23]. R.V. Guha, R. McCool, and E. Miller, "Semantic Search," *Proc. 12th Int'l Conf. World Wide Web* (WWW 03, May 2003, Budapest, Hungary).

[24]. S. Alexaki et al., "The ICS-FORTH RDFSuite: High-level Scalable Tools for the Semantic Web," *ERCIM NEWS*, news no. 51, Oct. 2002, http://www.ercim.org/publication/Ercim_News/enw51/alexaki.html.

[25]. S. Kokkelink and R. Schwänzl, *Expressing Qualified Dublin Core in RDF/ XML*, May 2002, http://dublincore.org/documents/dcq-rdf-xml/.

[26]. S. B. Palmer, RDF in HTML: Approaches, Jun. 2002, http://infomesh.net/2002/rdfinhtml/#hyperrdf

[27]. T.R. Payne, R. Singh, and K. Sycara, "Calendar Agents on the Semantic Web," *IEEE Intelligent Systems*, vol. 17, no. 3, May/June 2002, pp. 84-86.

[28]. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, May 2001, vol. 284, pp. 34-43, http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[29]. T. Berners-Lee, *Primer: Getting into RDF & Semantic Web using N3*, Apr. 2003, http://www.w3.org/2000/10/swap/Primer.html.

[30]. Y. Lafon and B. Boss, *Describing and retrieving photos using RDF and HTTP*, W3C Note, Apr. 2002, http://www.w3.org/TR/photo-rdf/.