# THE SEMANTIC WEB: AN EMERGING INFORMATION TECHNOLOGY

**Dr. Lissa F. Pollacia, Northwestern State University, pollacia@nsula.edu**
**Dr. Claude Simpson, Southeastern Oklahoma State University, csimpson@sosu.edu**

## ABSTRACT

*The purpose of this paper is to explain an emerging technology called the Semantic Web. This new technology is an attempt to add more meaning, or semantics, to web documents, thus permitting more efficient and meaningful information retrieval. The languages that define the structure of information on the Web were designed for human use, and not for computing machinery. Thus when retrieving information from the Web of today, humans must still function as a filter and process the enormous amount of information that is usually returned. The purpose of the Semantic Web is to provide a mechanism by which a computer, assisting the human doing work via the Web, will be able to comprehend the meaning, and thus the content, of Web documents. In addition to explaining the Semantic Web, this paper will present research accomplished on this project to date, focusing on XML (Extensible Markup Language) and Document Type Definition (DTD). XML permits a Web developer to create more descriptive tabs for objects on a web page, thus producing a better searching mechanism. In conclusion, the future of the Semantic Web will be discussed.*

**Keywords:** Semantics, XML, Web technology, Markup Language

## INTRODUCTION

The term *semantic* is derived from the Greek word for sign or signify. Today the term is commonly used in the computing world to mean "of or relating to *meaning*, often in language." [SWA 2001] The goal of the Semantic Web project is to develop language(s) for better defining the information in a Web document, thus enabling a computer to process this information efficiently, more like a human would do. Berners-Lee explains it best in his paper "Semantic Web Road Map" [1]:

> The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. One of the major obstacles to this has been the fact that most information on the Web is designed for human consumption, and… the structure of the data is not evident to a robot browsing the Web. Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine processable form.

## THE CURRENT SITUATION

Suppose you want to read a movie review of "The Lord of the Rings" in order to decide whether or not you want to see the movie. You conduct a search of the Web for "Lord of the Rings"

using Google, and get back over 1,710,000 hits in 0.05 seconds. (Even Google seems unsure of the exact number when there are so many hits to choose from!) [4]

Among the sites returned are those with movie merchandise, sites with information about the movie stars, fan club sites, guides to "Middle Earth", and even one that claims to tell the story using Legos, just to mention a few. Further down in the links are references to Muhammed Ali (the first "lord of the rings") and even men's gymnastics. You pour yourself another cup of coffee, and try to keep your eyes from glazing over at the prospect of trying to filter through this enormous maze of information.

Now you narrow down the search to Lord of the Rings *movie reviews*, thinking that should do it. Yes, now the number of links returned is just 319,000. Google had to work a little harder at this task and took 0.20 seconds this time before putting the ball squarely back in your court. Again, you must sift through the results. The problem is that the search engine cannot distinguish between the rambling writings of anyone who has seen the movie, and bona fide movie reviews. Which of these 319,000 sites can be "trusted" to contain a professional review? Hopefully, the solution to your quest will emerge in the first ten or twenty hits, before you give up and just decide to trust your friend's opinion and go see it. You've already spent more time than it would take to see the movie.

This illustrates the situation with searching the Web of today. Web content is designed for a person, i.e. a biological processor, to read and determine the meaning. Computers can efficiently parse the layout of the page, determining headers, fonts, and links, and to display that page according to the HTML formatting tags. Computers can conduct searches for keywords "movie, reviews, Lord, Rings, fantasy" and try to find those having the highest concentration of these keywords, thus hopefully returning those that are most relevant. But at this time, there is no way to determine what information a web document really contains.

## THE SEMANTIC WEB

The Semantic Web, as defined by Berners-Lee, is not a new and separate Web, but is an extension of the existing Web in which computers and people will be able to work cooperatively. It is the first attempt to weave semantics into the Web and extend the functionality of computers by enabling them to process and understand the data that they simply manipulate and display now. We want computers, acting as our agents, to be able to make inferences, answer questions, and help us choose between alternatives courses of action [2].

One of the fundamental properties of the Web is its lack of centralized control. A hyperlink can link any document to any other document on the Web. This has enabled a Web technology that does not distinguish between a hastily constructed student's paper and a polished professional thesis, between commercial and non-commercial material, between languages, culture, etc. Berners-Lee states that another axis along which Web material greatly varies is the one that has human consumption on one end, and computer/machine processing on the other. The Web has been progressing mostly at the human consumption end of the scale; i.e. Web documents are a medium primarily for people rather than for data that can be processed automatically [2].

The Semantic Web attempts to change this, to some degree. In order to do this, machines must have access to structured data that better describes the Web documents, i.e. more expressive data. Thus the Semantic Web must provide a new language, one that can express both the structure of the data *along with* rules for reasoning about the data, i.e. rules of inference. Three important languages/technologies already exist that can fill this role: the *eXtensible Markup Language (XML),* and *Document Type Definitions (DTD's)*. These are described in the following sections.

## THE EXTENSIBLE MARKUP LANGUAGE (XML)

An important language that will be used to develop the Semantic Web is known as the eXtensible Markup Language (XML). A working group of the World Wide Web Consortium (W3C) completed the development of XML in 1998 [3], [7]. The original goal of this W3C working group was to bring the power of SGML (Standard Generalized Markup Language), also known as the parent language of HTML, to the Web. The result was XML, a faster, trimmed-down version of SGML specifically designed for Web applications. Since its introduction in 1998, the use of XML has grown tremendously and is being utilized to develop many different applications on the web today.

Basically, XML permits Web authors to create their own tags to describe the data contained in a Web document. For example, an XML document containing a recipe might have the tags <Recipe>, <Ingredients>, <CookingInstructions>, and <ServingSuggestions>. There might be further refinement of the data, such as <Step> … </Step> text within the <CookingInstructions> section. A human reader of the document would be able to tell what kind of data each tag is designed to describe. Thus XML allows the Web author to add more refined structure to their documents. Here is a simple XML document to display one of the author's favorite recipes:

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE Recipe SYSTEM "Recipe.dtd">   <!- document type definition ->
<Recipe cook="Lissa's Favorite Recipes">
   <Title>Lissa's Pork Roast</Title>
   <Category name="Pork Entrée" />

   <Ingredients>
        <Item>1 two to three pound pork loin</Item>
        <Item>3 onions, thinly slices</Item>
        <Item>olive oil</Item>
        <Item>salt and pepper</Item>
        <Item>garlic powder</Item>
  </Ingredients>

  <CookingInstructions>
        <Step>Pour olive oil into a large skillet or Dutch oven.</Step>
        <Step>Season the pork loin on all sides with salt, pepper, and garlic.</Step>
        <Step>Brown pork loin in heated skillet, turning to brown all sides.</Step>
        <Step>Add about 1/2 cup of water and turn down fire to medium.</Step>
        <Step>Place slices of onion on top of pork and all around sides.</Step>
```

```
        <Step>Cook 3 – 4 hours on med/low heat.  May have to add more water.</Step>
  </CookingInstructions>

  <ServingSuggestions>
        Remove pork from pot, slice, and place on serving plate.  Pour the gravy and onions on
top of the meat and serve immediately.  Goes well with a rice side dish.
    </ServingSuggestions>
</Recipe>
```

As can be seen above, XML makes it easier to encode some of the semantics, at least to a human reader, into the document itself. This makes it easier for search engines, using XML tag contents and attributes, to find more specific information. Search engines now rely on document titles and headings encoded in HTML. Using XML, search engines will search the tags that are embedded throughout the document.

## DOCUMENT TYPE DEFINITIONS

XML permits Web authors to define markup tags like <Ingredients> that carry more meaning to us. A computer can build a tree structure with the elements of the document and parse it to determine if it is *well-formed*. However, to formally define the structure of the XML document, a *Document Type Definition*, or DTD, is recommended. A DTD explicitly defines the rules of what elements can be used in certain markups and how those elements should be sequenced. In other words, the DTD consists of the syntax specifications that spell out the grammatical rules of the markup language. It is a type of metalanguage, similar to *the Backus-Naur Form*, or BNF, that is a commonly used notation for defining programming languages [5].

A DTD is usually defined in a separate file and will usually be given a declaration within the XML document, such as:

```
<!DOCTYPE  Recipe  SYSTEM  "Recipe.dtd">
```

In this example, Recipe.dtd is a relative *Uniform Resource Identifier (URI)* that specifies where the DTD is located. A URL is a type of URI. The DTD then contains a specification of the elements and their attributes. For example, the recipe DTD might contain the following:

```
<!DOCTYPE Recipe [
<!ELEMENT Recipe (Title, Category, Ingredients, CookingInstructions, ServingInstructions?)
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Category (#PCDATA)>
<!ELEMENT Ingredients (Item+)>
<!ELEMENT Item (#PCDATA)>
<!ELEMENT CookingInstructions (#PCDATA)>
<!ELEMENT ServingInstructions (#PCDATA)>
]>
```

This DTD specifies the root element, Recipe, is composed of a Title, followed by a Category, followed by Ingredients (one element, which is composed of one or more Item elements), then CookingInstructions, and finally ServingInstructions, which are optional. PCDATA stands for *parsed character data*, or text.

A validating parser will then compare the XML document to the declared DTD. If the content of the document is valid, the document is then passed to an application, such as a browser. XML uses style sheets for formatting the document and controlling how it will be displayed. This separation of a document's content with the display formatting enables the same document to be displayed in different ways, depending upon the device. For example, the recipe would be displayed differently on a hand-held device than it would be on a desktop computer.

## FUTURE OF THE SEMANTIC WEB

There is no doubt that the language of the Web is undergoing change, i.e. that HTML will be replaced with XML or possibly XHTML, which is a form of XML that includes the syntax of HTML. There is also some work underway to develop ontologies. Ontologies are a way to give differing labels some common meaning, even crossing international borders. Thus the terms *zip*, *zipcode*, and *mailingcode* will all have the same meaning. Computers need ontologies to formalize agreement on the meaning of XML tags created in different environments or by different user groups.

Another work in progress is W3C's *Resource Description Framework* [8], which is an XML text format that supports resource description and metadata applications. The purpose of RDF is to integrate applications and computer "agents", acting on your behalf, into the one Semantic Web. With RDF, comes the vision that something can be done to solve the chaotic and overwhelming web searching that exists today. This is the fundamental vision of the Semantic Web.

## REFERENCES

[1] Berners-Lee, T. "Semantic Web Road Map", World Wide Web Consortium, Sept. 1998. www.w3.org/DesignIssues/Semantic.html

[2] Berners-Lee, T. "The Semantic Web", *Scientific American*, May 1, 2001. www.sciam.com/2001/0501issue/0501berners-lee.html

[3] Bosak, J. and Bray, T. "XML and the Second-Generation Web", *Scientific American*, May 1, 1999. http://www.sciam.com/1999/0599issue/0599bosak.html

[4] Long, P. D. "Weaving Semantics Into the Web", *Syllabus*, vol. 15, no. 7, Syllabus Press, Feb. 2002, 8-10.

[5] Sebesta, R. W. *Concepts of Programming Languages*. Addison-Wesley, Reading, MA, 1999.

[6] Semantic Web Activity Statement, contact: Eric Miller, W3C Semantic Web Activity Lead, November, 2001. http://www.w3.org/2001/sw/Activity

[7] World Wide Web Consortium, Extensible Markup Language (XML), Reference document, 1997. www.w3.org/XML/

[8] World Wide Web Consortium, Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation, www.w3.org/TR/REC-rdf-syntax/