Security Challenges in Implementing Semantic Web- Unifying Logic

Nana Kwame Gyamfi
Department of Computer Science, P.O. box 854 Kumasi-Ashanti

Abstract - The Web is entering a new phase of evolution. There has been much debate recently about what to call this new phase. Some would prefer to not name it at all, while others suggest continuing to call it “Web 2.0”. Some others suggested naming this third-generation of the Web, “Web 3.0”. This suggestion has led to quite a bit of debate within the industry. However, this new phase of evolution has quite a different focus from what Web 2.0 has come to mean. The most controversial feature of the Web 3.0 is the Semantic Web. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF). In this paper, we describe the Semantic Web, which provides an intermediate solution.

Keywords - Semantic web, Security Challenges, Unifying logic, RDF.

1. Introduction

The World Wide Web (abbreviated as WWW or W3, commonly known as the web) is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them via hyperlinks. Since the first Web 2.0 conference hosted by O’Reilly Media and MediaLive in 2004, the term “Web 2.0” has become very popular. Web 2.0 websites allow users to do more than just retrieve information. Instead of merely ‘reading’, a user is invited to ‘write’ as well, or contribute to the content available to everyone in a user friendly way.

Major features of Web 2.0 include social networking sites, user created web sites, self-publishing platforms, tagging, and social bookmarking. Users can provide the data that is on a Web 2.0 site and exercise some control over that data. The key features of Web 2.0 include:

• Folksonomy- free classification of information; allows users to collectively classify and find information (e.g. Tagging).
• Rich User Experience- dynamic content; responsive to user input
• User as a Contributor- information flows two ways between site owner and site user by means of evaluation, review, and commenting
• Long tail services offered on demand basis; profit is realized through monthly service subscriptions more than one-time purchases of goods over the network
• User Participation- site users add content for others to see
• Basic Trust- contributions are available for the world to use, reuse, or re-purpose
• Dispersion- content delivery uses multiple channels, digital resources and services are sought more than physical goods
• Mass Participation

These features empowered by Web 2.0 technologies such as Ajax and JavaScript frameworks such as YUI Library, Dojo Toolkit, MooTools, jQuery, Ext JS and Prototype JavaScript Framework are making up the world Wide Web as we know it today.

However, while the innovations and practices of Web 2.0 will continue to develop, they are not the final step in the evolution of the Web. In fact, there is a lot more in store for the Web. We are starting to witness the convergence of several growing technology trends that are outside the scope of what Web 2.0 has come to mean. These trends have been gestating for a decade and will soon reach a tipping point. At this juncture the Web 3.0 will start.

In addition to the classic “Web of documents” W3C is helping to build a technology stack to support a “Web of data,” the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term “Semantic Web” refers to W3C’s vision of 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. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS. It has been described in rather different ways: as a utopic vision, as a web of data, or merely as a natural paradigm shift in our daily use of the Web. Is the Semantic Web really utopic?

1.1 Introduction

Today's Web has terabytes of information available to humans, but hidden from computers. It is a paradox that information is stuck inside HTML pages, formatted in esoteric ways that are difficult for machines to process. In this paper, we describe the Semantic Web, which provides an intermediate solution.

2. The Web 3.0

The Web 3.0, is likely to be a pre-cursor of the real semantic web, is going to make our computers more intelligent by making them capable of processing web data by themselves. What we mean by 'Web 3.0' is that major web sites are going to be transformed into web services - and will effectively expose their information to the world.

There are actually several major technology trends that are about to reach a new level of maturity at the same time. The simultaneous maturity of these trends is mutually reinforcing, and collectively they will drive the third-generation Web. From this broader perspective, Web 3.0 might be defined as a third-generation of the Web enabled by the convergence of several key emerging technology trends:

- Ubiquitous Connectivity
  - Broadband adoption
  - Mobile Internet access
  - Mobile devices
- Network Computing
  - Software-as-a-service business models
  - Web services interoperability
  - Distributed computing (P2P, grid computing, hosted “cloud computing” server farms such as Amazon S3)
- Open Technologies
  - Open APIs and protocols
  - Open data formats
  - Open-source software platforms
  - Open data (Creative Commons, Open Data License, etc.)
- Open Identity
  - Open reputation
  - Portable identity and personal data (for example, the ability to port your user account and search history from one service to another)
  - The Intelligent Web
  - Semantic Web technologies (RDF, OWL, SWRL, SPARQL, Semantic application platforms, and statement-based datastores such as triplestores, tuplestores and associative databases)
  - Distributed databases or what I call “The World Wide Database” (wide-area distributed database interoperability enabled by Semantic Web technologies)
  - Intelligent applications (natural language processing, machine learning, machine reasoning, autonomous agents)

Figure 1 - Evolution of the Web (source: Radar Networks & Nova Spivack 2007)
Web 3.0 will be more connected, open, and intelligent, with semantic Web technologies, distributed databases, natural language processing, machine learning, machine reasoning, and autonomous agents.

3. The Semantic Web

a. Definition

The Semantic Web is a collaborative movement led by international standards body the World Wide Web Consortium (W3C). The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a “web of data”.

Tim Berners-Lee’s original vision clearly involved more than retrieving Hypertext Markup Language (HTML) pages from Web servers. A more understandable Web should present relations between information items like “includes,” “describes,” and “wrote”. Unfortunately, these relationships between resources are not currently captured on the Web. The technology to capture such relationships is called the Resource Description Framework (RDF).

So, how do we create a web of data that machines can process? The first step is a paradigm shift in the way we think about data. Historically, data has been locked away in proprietary applications. Data was seen as secondary to processing the data. This incorrect attitude gave rise to the expression “garbage in, garbage out,” or GIGO. GIGO basically reveals the flaw in the original argument by establishing the dependency between processing and data. In other words, useful software is wholly dependent on good data. Computing professionals began to realize that data was important, and it must be verified and protected.

Programming languages began to acquire object-oriented facilities that internally made data first-class citizens. However, this “data as king” approach was kept internal to applications so that vendors could keep data proprietary to their applications for competitive reasons. With the Web, Extensible Markup Language (XML), and now the emerging Semantic

Web, the shift of power is moving from applications to data. This also gives us the key to understanding the Semantic Web. The path to machine-processable data is to make the data smarter.

b. Purpose

The HTML language as we know it today has several limitations. Many files on a typical computer can be loosely divided into human readable documents and machine readable data. Documents like mail messages, reports, and brochures are read by humans. Data, like calendars, address books, playlists, and spreadsheets are presented using an application program which lets them be viewed, searched and combined.

Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML), a markup convention that is used for coding a body of text interspersed with multimedia objects such as images and interactive forms. Metadata tags provide a method by which computers can categories the content of web pages, for example:

```html
<meta name="keywords" content="computing, computer studies, computer" />
<meta name="description" content="Cheap widgets for sale" />
<meta name="author" content="John Doe" />
```

With HTML and a tool to render it (perhaps web browser software, perhaps another 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 ‘Widget Superstore’”, but there is no capability within the HTML itself to assert
unambiguously that, for example, item number X586172 is an Acme Gizmo with a retail price of €199, or that it is a consumer product. Rather, HTML can only say that the span of text "X586172" is something that should be positioned near "Acme Gizmo" and "€199", etc. There is no way to say "this is a catalog" or even to establish that "Acme Gizmo" is a kind of title or that "€199" 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.

Semantic HTML refers to the traditional HTML practice of markup following intention, rather than specifying layout details directly. For example, the use of <em> denoting "emphasis" rather than <i>, which specifies italics. Layout details are left up to the browser, in combination with Cascading Style Sheets. But this practice falls short of specifying the semantics of objects such as items for sale or prices.

Microformats extend HTML syntax to create machine-readable semantic markup about objects including people, organizations, events and products. Similar initiatives include RDFa, Microdata and Schema.org.

The main purpose of the Semantic Web is driving the evolution of the current Web by enabling users to find, share, and combine information more easily. Humans are capable of using the Web to carry out tasks such as finding the Estonian translation for "twelve months", reserving a library book, and searching for the lowest price for a DVD. However, machines cannot accomplish all of these tasks without human direction, because web pages are designed to be read by people, not machines. The semantic web is a vision of information that can be readily interpreted by machines, so machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web.

The Semantic Web, as originally envisioned, is a system that enables machines to "understand" and respond to complex human requests based on their meaning. Such an "understanding" requires that the relevant information sources be semantically structured.

The Semantic Web is regarded as an integrator across different content, information applications and systems. It has applications in publishing, blogging, and many other areas. Often the terms "semantics", "metadata", "ontologies" and "Semantic Web" are used inconsistently. In particular, these terms are used as everyday terminology by researchers and practitioners, spanning a vast landscape of different fields, technologies, concepts and application areas. Furthermore, there is confusion with regard to the current status of the enabling technologies envisioned to realize the Semantic Web. The architectural model proposed by Tim Berners-Lee is used as basis to present a status model that reflects current and emerging technologies.

3. Implementation

i. Overview

The Semantic Web involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts.

These technologies are combined in order to provide descriptions that supplement or replace the content of Web documents. Thus, content may manifest itself as descriptive data stored in Web-accessible databases, or as markup within documents (particularly, in Extensible HTML (XHTML) interspersed with XML, or, more often, purely in XML, with layout or rendering cues stored separately). The machine-readable descriptions enable content managers to add meaning to the content, i.e., to describe the structure of the knowledge we have about that content. In this way, a machine can process knowledge itself, instead of text, using processes similar to human deductive reasoning and inference, thereby obtaining more meaningful results and helping computers to perform automated information gathering and research.

Tim Berners-Lee calls the resulting network of Linked Data the Giant Global Graph, in contrast to the HTML-based World Wide Web. Berners-Lee posits that if the past was document sharing, the future is data sharing. His answer to the question of "how" provides three points of instruction. One, a URL should point to the data. Two, anyone accessing the URL should get data back. Three, relationships in the data should point to additional URLs with data.

ii. RDF (Resource Description Framework)

RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without
requiring all the data consumers to be changed.

RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

An example of a tag that would be used in a non-semantic web page:

```html
<Item>blog</Item>
```

Encoding similar information in a semantic web page might look like this:

```html
<Item rdf:about="http://example.org/semantic-web/">Semantic Web</Item>
```

This linking structure forms a directed, labeled graph, where the edges represent the named link between two resources, represented by the graph nodes. This graph view is the easiest possible mental model for RDF and is often used in easy-to-understand visual explanations.

The RDF data model is similar to classic conceptual modeling approaches such as entity–relationship or class diagrams, as it is based upon the idea of making statements about resources (in particular web resources) in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. For example, one way to represent the notion "The sky has the color blue" in RDF is as the triple: a subject denoting "the sky", a predicate denoting "has the color", and an object denoting "blue". Therefore RDF swaps object for subject that would be used in the classical notation of an entity–attribute–value model within object-oriented design; object (sky), attribute (color) and value (blue).

RDF's simple data model and ability to model disparate, abstract concepts has also led to its increasing use in knowledge management applications unrelated to Semantic Web activity. A collection of RDF statements intrinsically represents a labeled, directed multi-graph. As such, an RDF-based data model is more naturally suited to certain kinds of knowledge representation than the relational model and other ontological models. However, in practice, RDF data is often persisted in relational database or native representations also called Triplestores, or Quad stores if context (i.e. the named graph) is also persisted for each RDF triple.

The subject of an RDF statement is either a uniform resource identifier (URI) or a blank node, both of which denote resources. Resources indicated by blank nodes are called anonymous resources. They are not directly identifiable from the RDF statement. The predicate is a URI which also indicates a resource, representing a relationship. The object is a URI, blank node or a Unicode string literal. The predominant query language for RDF graphs is SPARQL. SPARQL is an SQL-like language, and a recommendation of the W3C as of January 15, 2008.

iii. OWL (Web Ontology Language)

The OWL Web Ontology Language is an international standard for encoding and exchanging ontologies and is designed to support the Semantic Web. In order for information from different sources to be integrated, there needs to be a shared understanding of the relevant domain. Knowledge representation formalisms provide structures for organizing this knowledge, but provide no mechanisms for sharing it. Ontologies provide a common vocabulary to support the sharing and reuse of knowledge.

The Web Ontology Language (OWL) is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases. The languages are characterized by formal semantics and RDF/XML-based serializations for the Semantic Web. Therefore, syntactically, an OWL ontology is a valid RDF document and as such also a well-formed XML document.

iv. XML (eXtensible Markup Language)

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications, all free open standards.

The design goals of XML emphasize simplicity, generality, and usability over the Internet. It is a textual data format with strong support via Unicode for the
languages of the world. Although the design of XML focuses on documents, it is widely used for the representation of arbitrary data structures, for example in web services.

Many application programming interfaces (APIs) have been developed to aid software developers with processing XML data, and several schema systems exist to aid in the definition of XML-based languages. The technologies that XML is built upon are Unicode characters and Uniform Resource Identifiers (URIs). The Unicode characters allow XML to be authored using international characters. URIs are used as unique identifiers for concepts in the Semantic Web.

v. An example of implementation

The Amazon E-Commerce API - open access to Amazon's catalog. Amazon has revealed these last year a visionary WebOS strategy. The Seattle web giant is reinventing itself by exposing its own infrastructure via a set of elegant APIs. One of the first web services opened up by Amazon was the E-Commerce service. This service opens access to the majority of items in Amazon's product catalog. The API is quite rich, allowing manipulation of users, wish lists and shopping carts. However its essence is the ability to lookup Amazon's products.

Why has Amazon offered this service completely free? Because most applications built on top of this service drive traffic back to Amazon (each item returned by the service contains the Amazon URL). In other words, with the E-Commerce service Amazon enabled others to build ways to access Amazon's inventory. As a result many companies have come up with creative ways of leveraging Amazon's information. Here is an illustration of the net effect of Amazon was the E-Commerce service:

So bringing together Open APIs (like the Amazon E-Commerce service), gives us a way to treat any web site as a web service that exposes its information. The information, or to be more exact the data, becomes open. In turn, this enables software to take advantage of this information collectively. With that, the Web truly becomes a database that can be queried and remixed. This paper is intending to provide a step by step process of building a remote accessible application implementing several ontologies.

vi. The Unifying Logic

The unifying logic described into the Semantic Web stack can be implemented in several ways. This paper is intending to provide a standardization, a general implementation method that would cover several aspects of the semantic web. Our work on a common framework is directed to provide the Semantic Web community a foundational tool which can be the basis for the discussion towards a common rule language for the Semantic Web with an agreed upon semantics.

4. Research Methodology

4.1. Overview

Our main objective is to to present the best practices in the implementation of Semantic Web, our research will be performed following the Case Study research methodology. It is a research methodology which allow a high level of discoverability and represent ability.

4.2. The Case Study Design

a. Definition

A case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between the phenomenon and context are not clearly evident. The case study method has seen extensive application in Information Systems (IS). The case study approach refers to a group of methods which emphasize qualitative analysis. Data are collected from a small number of organizations through methods such as participant-observation, in-depth interviews, and longitudinal studies. The case study approach seeks to understand the problem being investigated. It provides the opportunity to ask penetrating questions and to capture the richness of organizational behavior, but the conclusions drawn may be specific to the particular organizations studied and may not be generalizable. Yin (1994) presented at least
four applications for a case study model:

- To explain complex causal links in real-life interventions
- To describe the real-life context in which the intervention has occurred
- To describe the intervention itself
- To explore those situations in which the intervention being evaluated has no clear set of outcomes.

Information technologies involve all four of the above categories, but this study will only report on the last two. Since the Levy (1988) case study of the University of Arizona, there has been very little literature relating to the pace of acquisition of information technology at institutions of higher education. For this reason, Levy (1988) conducted a case study after consulting with experts in the field and with senior case researchers. Their recommendation was to conduct an in-depth study of the institution using the case methodology. This study replicates and extends that study and thereby adds to the body of knowledge on the nature of information technology acquisition at universities.

b. Methodology

Yin (1984) has divided the methodology of Case Study Research into 4 main stages. Each stage of the methodology will consist of a discussion of procedures recommended in the literature, followed by a discussion of the application of those procedures in the proposed study:

- Design the case study protocol: The protocol should include the following sections:
  - An overview of the case study project - this will include project objectives, case study issues, and presentations about the topic under study

<table>
<thead>
<tr>
<th>Source of Evidence</th>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documentation</td>
<td>stable - repeated review</td>
<td>retrievability - difficult</td>
</tr>
<tr>
<td></td>
<td>unobtrusive - exist prior to case study</td>
<td>biased selectivity</td>
</tr>
<tr>
<td></td>
<td>Exact - names etc.</td>
<td>reporting bias - reflects</td>
</tr>
<tr>
<td></td>
<td>broad coverage - extended time span</td>
<td>author bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>access - may be blocked</td>
</tr>
<tr>
<td>Archival Records</td>
<td>Same as above</td>
<td>Same as above</td>
</tr>
<tr>
<td></td>
<td>precise and quantitative</td>
<td>privacy might inhibit access</td>
</tr>
<tr>
<td>Interviews</td>
<td>targeted - focuses on case study topic</td>
<td>bias due to poor questions</td>
</tr>
<tr>
<td></td>
<td>insightful - provides perceived causal inferences</td>
<td>response bias</td>
</tr>
<tr>
<td></td>
<td></td>
<td>incomplete recollection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reflexivity - interviewee</td>
</tr>
<tr>
<td></td>
<td></td>
<td>expresses what interviewer wants to hear</td>
</tr>
<tr>
<td>Direct Observation</td>
<td>reality - covers events in real time</td>
<td>time-consuming</td>
</tr>
<tr>
<td></td>
<td>contextual - covers event context</td>
<td>selectivity - might miss facts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reflexivity - observer’s presence might cause change</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cost - observers need time</td>
</tr>
<tr>
<td>Participant</td>
<td>Same as above</td>
<td>Same as above</td>
</tr>
<tr>
<td>Observation</td>
<td>insightful into interpersonal behavior</td>
<td>bias due to investigator's actions</td>
</tr>
<tr>
<td>Physical Artifacts</td>
<td>insightful into cultural features</td>
<td>selectivity</td>
</tr>
<tr>
<td></td>
<td>insightful into technical operations</td>
<td>availability</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 - Strengths and Weaknesses of evidences

Field procedures - reminders about procedures, credentials for access to data sources, location of those
My investigations have to cover a large amount of data types in order to explain in detail the implementation methods. These data types will be diverse including standard data types (int, char, String…) and Semantic data types (length, price, area…). All the simulations are going to be proceeded on a single local server.

6. Conclusion

The high-level goal of the researcher was to help tackling the information overload (on the web). This paper has given an overview of the developing Semantic Web infrastructure, showed how this relates to typical hypermedia research topics and given comprehensive pointers to the relevant literature. The introduction of the Web, as a highly distributed, but relatively simple, hypermedia system has also influenced the character of hypermedia research. The existence of XML and RDF, along with developments such as RDF Schema, provide the impetus for realizing the Semantic Web. During these early stages of its development, I want to ensure that the many hypertext lessons learned in the past will not be lost, and that future research tackles the most urgent issues of the Semantic Web.

As a potential outcome from this study we can expect a list of best practices, a generalized unifying language and a set of tools useful for an optimal implementation of a Semantic web application.

Acknowledgment

I will like to acknowledge all members of staff in Graduate studies (Computer Science) of University of Ghana and all MSc/Mphil/PhD class members.
References


Author

Nana Kwame Gyamfi [BSc ICT, MSc I.T, Ethical Certified Security Specialist, ITIL Foundation, MCT, MCSE, CompTIA A+, N+, Security +, CCNA, Oracle Database 11g (OCA, OCP, RAC, Security, Performance Tuning, Solaris 10 Associate and Professional)]
Kumasi Polytechnic
Department of Computer Science
P.O. box 854 Kumasi-Ashanti
emeritusbrains@gmail.com
Member, IACSIT
An experienced I.T professional with four (5) years hands-on work experience in the provision of consulting services to corporate bodies such as B&D credit Bureau (Accra), Adansi North Education (Kumasi). My key strength and consulting work experiences have been related to the following: Hardware and Networking, Systems Administration, IT Security Policy, Documentation, Software Development, System Analysis and Design, Data Recovery, Website Development, ICT Training and IT services Management, Oracle Database 11g Security Certified Implementation Specialist and Oracle Solaris 10 System Administrator Certified Professional I. Currently, I provide teaching and system administration services to Kumasi Polytechnic on full time and SMU on part-time. I hold an MSc in I.T and BSc in Information and Communication Technology. In addition, I have acquired international professional certificate: My current research concentration is on Information Security paths.