ON THE SEMANTIC WEB AND ITS WEB OF TRUST

Wesley Brandi
Department of Computer Science, Rand Afrikaans University
PO Box 524, Auckland Park, Johannesburg, 2006 South Africa
wb@eclab.rau.ac.za

Abstract In a recent paper by Berners-Lee et al [2] the World Wide Web of the future is proposed. It is a Web where data is defined in a manner that allows machines to "understand" the content that is available on the Web. No longer will the content of documents be understood by humans alone. The Web of which we speak promises a plethora of features and functionality not yet possible using the World Wide Web of today. Berners-Lee et al speak of giving meaning to data in the Web of the future, the Semantic Web.

The security model of the Semantic Web per se is dependent on a Web of Trust, as proposed by Swartz et al [13]. This Web of Trust is similar to the Web of Trust implemented by PGP [8] in that A (who trusts B) will trust C if B trusts C. In this paper we examine this Web of Trust and show that it may not be adequate when one considers the nature of the Web today.

Keywords: Semantic web, web of trust, security

1. INTRODUCTION

In 1989 Tim Berners-Lee proposed a non-linear text system which would allow for any hypertext document in the system to reference any other hypertext document in the system [1]. References could be made in a completely autonomous fashion. It was not necessary for the owner of the referenced document to be aware that the document was referenced, nor was it necessary that the referenced document even existed.

The proposed system would also allow for the following:

- Documents could reference across remote networks.
- Users of different platforms (Unix, VM/CMS, Macintosh, VAX/VMS) could access the same data.
- There was no need to establish a central point of control. Existing systems could therefore be linked together without the need of for any external coordination.

Today, we know this system as the World Wide Web. In allowing users of the Web to manage links and documents in an autonomous fashion, the Web has demonstrated exponential growth in a very short time. Unfortunately, the inherent flexibility of the Web has come at a price. There are approximately 300 million users of the Web and currently they have access to more than 3 billion documents. This enormous amount of information has made finding anything on the Web a sisyphean endeavour.

Search engines trawl the Web and employ sophisticated search algorithms upon the trawled data. The search engines then provide users with the service of helping them to find what they are looking for on the Web. Although search engine technology has improved tremendously when one compares the search engines of 10 years ago to the search engines of today, the task of finding exactly what one is looking for is still tedious and difficult.

The primary reason for this is not only due to the magnitude of available data on the Web but more so because the data represented on the Web is in natural language. A search for "how many banks are there in South Africa?" will most likely not return the result one is looking for. A search engine will break up the search into keywords and search for the most relevant page pertaining to the keywords.

The Semantic Web offers a solution to this problem. Consider the implications of machines that could "understand" the data that they are processing. Note that we do not use the term understand in the same way one would use it when referring to humans understanding

certain principles. By machines understanding data, we simply refer to machines having access to well structured, well defined data. Section 2.2 discusses how, with the use of agents, this query is addressed using the Semantic Web.

In this paper we introduce the Semantic Web as proposed by Berners-Lee et al [2]. We then discuss the proposed model of trust for the Semantic Web and identify certain pitfalls that may arise when it is implemented. In particular, we intend to show that the model of trust will fail when implemented in a Web that is dominated by businesses.

This paper is structured as follows: Section 2 will serve as an introduction to the Semantic Web and its components, including the proposed Web of Trust. Analysing the pitfalls of the trust model requires that we first provide a better understanding of the environment the Semantic Web will be functioning in. Therefore, in Section 3 we discusses Businesses and their growing participation on the Web. Section 4 makes use of two examples to demonstrate how the Web of Trust will succeed and fail when implemented in a Web dominated by business. This paper is then concluded in section 5.

2. BACKGROUND

The Semantic Web is not a new Web that will be created separate from the current Web. Instead, it is proposed as an extension to the Web we know today. The Semantic Web's primary objective is to give machines (or software agents) the ability to process well defined data in a far more effective manner. Having access to well defined data will allow software agents to perform complex tasks on enormous amounts of data.

Everything on the Web can be identified by means of an identifier commonly known as a Uniform Resource Identifier. A typical example of a URI is the Uniform Resource Locator (URL). The URL *http://www.rau.ac.za* is said to be an identifier (URI) for the Rand Afrikaans University (RAU).

If Bobby wishes to say something along the lines of "Bobby likes RAU" on his Web Site, using the standard markup language of the Web today, it may look something like the following:

```
<html>
  Bobby likes RAU
</html>
```

Machines can not really process data of this nature in an effective manner, simply because nothing is really known about the data. The data is not well defined.

The Extensible Markup Language (XML) remedies this problem by allowing users to format their data to enhance its meaning. Once formatted, the above statement in XML may look as follows:

```
<student>Bobby</student> likes <university>RAU</university>
```

There is nothing stopping users creating their own tags (with the same names) to portray different meanings. User A may define the tag HOME to be used when making reference to the web site that is referenced when clicking the home button on his browser. User B may define the HOME tag to indicate his place of residence.

To avoid these conflicts users have the option of defining XML Namespaces.

2.1. RESOURCE DESCRIPTION FRAMEWORK (RDF) AND ONTOLOGIES

The proposed language of the Semantic Web (Resource Description Framework) takes this further. Every RDF statement is almost entirely made up of URIs and is broken up into three parts: subject, predicate and object. "Bobby likes RAU" already has the right structure: "Bobby" is the subject, "likes" the predicate and "RAU" the object. Once formatted to URIs they would make up a simple RDF statement:

```
<http://www.rau.ac.za/students/bobby>
<http://definition.of.likes.org.za/> <http://www.rau.ac.za>
```

The real power of the Semantic Web is realised upon combining RDF with Ontologies. Ontologies essentially define and describe relationships among terms. Berners-Lee et al [2] use the following example:

An ontology may express the rule "If a city code is associated with a state code, and an address uses the city code, then that address has the associated state code." A program could then readily deduce, for instance, that a Cornell University address, being in Ithaca, must be in New York State, which is in the U.S., and therefore should be formatted to U.S. standards.

With reference to the example earlier, a software agent may deduce that since Bobby is a student and has a Web Site on RAU's student Web Server, Bobby is obviously a student at RAU.

For a thorough discussion of RDF and Ontologies readers referred to [3] and [13].

2.2. THE SEMANTIC WEB'S WEB OF TRUST

A problem with the Semantic Web is that "anything can say anything about anything." The Semantic Web will prove far more efficient than the current Web in handling queries the likes of "how many banks are there in South Africa?". Let us suppose that we now make use of an agent to help us answer this query using the Semantic Web. The agent knows that it must find the number of banking institutions (online or not) that are currently open for business in South Africa (a country on the continent of Africa).

In conducting its search it comes across a Web Site of someone who has already conducted exactly the same search. This person concludes that there are N amount of banks in South Africa. Does the agent stop its search and use these results as its answer? The answer depends on the level of trust associated with the person who made the statement.

The agent will consult its local Web of Trust. If there is an entry in the Web of Trust regarding this person that states he is trusted completely when making any statements at all, then the agent will conclude its search using his results.

If there is no definite entry regarding this person it will ask the trusted parties in its Web of Trust if they have any level of trust associated with the person who has made this statement. Degrees of trust can be associated with this party depending on how many levels of trusted parties are queried and to what extent the trust between them lies. If there is little to no trust between the agent and the party making the statement, the agent may decide only to notify the user that the Web Site of the statement may be worth looking through at a later stage.

3. BUSINESSES ON THE WEB

In this section we intend to show that businesses are becoming increasingly more active on the Web. We define a business and analyse results provided by NetCraft [11] that clearly show businesses are indeed dominating the Web.

We define a business as an entity that produces goods or services to meet people's needs. It is widely accepted that the primary objective of a business is to make a profit. Secondary objectives include expansion of the business so as to increase market share and share holder revenue.

Cronje et al [6] state that "the most important characteristic of the business world in developed countries of the West and Asia is the freedom individuals have to establish or terminate a business, to buy and sell shares in businesses and to produce, within limits, any product or service

the market requires." The inherent nature of the Web today is that it is extremely flexible. Other than consuming minimal resources to start a Web Site, maintenance of a Web Site can be done at any time from anywhere in the world. What better platform could one hope for to establish, promote and conduct business in a fashion as described by Cronje et al?

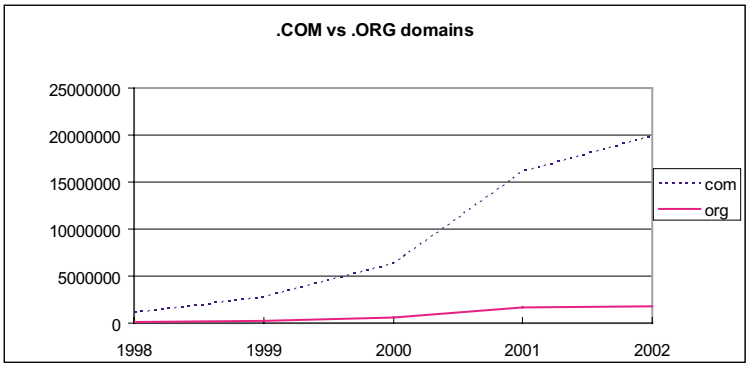

Figure 1 Depiction of the number of .COM domains vs .ORG domains on the Web from 1998 to 2002.

Figure 1 depicts the number of .COM domains vs .ORG domains on the Web from 1998 to 2002. The figures provided are the result of a number of surveys conducted by Netcraft [11]. In this figure it is clear that over several years the .COM domain has been far more popular than the .ORG domain. Comer [7] defines a .COM domain as a domain that is assigned to a commercial organisation. Organisations that do not represent governments, educational institutions, military groups or major network support centers may be assigned the .ORG domain.

If we use the number of registered high level domains on the Web to determine what the current nature of the Web may be, it is obvious from figure 1 that the Web is dominated by commercial institutions (businesses). Note that in referring to the nature of the Web we are particularly interested in the type of services being offered on publicly accessible domains rather than those offered by the likes of dial-up users on their private Web Sites.

4. BUSINESSES AND THE WEB OF TRUST

Having determined that the Web is being used on a large scale for business purposes, we now consider the following:

Big Bank has had much success in the years it has provided quality service to its customers. Inevitably, Big Bank has moved its services online. Although transparent to users of Big Bank's online services, Big Bank is always compliant with the most recent Semantic Web standards. Users of Big Bank's online services tend to agree that it is of a high standard, extremely efficient and easy to use. Big Bank is concerned over various security issues when banking online (as most banks are) and therefore only allow users to use the service over secure protocols. This requires that users trust Big Bank, and they in doing so, Big Bank becomes a part of the users Web of Trust.

4.1. THE WEB OF TRUST SUCCEEDS

User X of Big Bank ventures off to a site on the Web. There is a statement on the site by author Y that Big Bank steals money from the poor. In assessing the validity of the statement, or rather the weight that the author of the statement holds, the user's software (or agent) queries its local Web of Trust. The software finds that although there is no entry specifically for author Y in its local Web of Trust, Big Bank does have an entry. Big Bank knows of author Y and his slanderous comments and since Big Bank does not steal from the poor, Big Bank does not trust author Y.

Since thousands of people trust Big Bank, and since user X trusts Big Bank (especially with his money), the software determines that it makes sense to trust Big Bank in determining that whatever author Y says, particularly the statement in question, is false. User X therefore distrusts author Y.

Cases such as these are common and practically do make sense. When anyone can publish anything (good or bad) on the Web about any company or any person, the Web of Trust will serve as an invaluable tool in determining who to trust.

4.2. THE WEB OF TRUST FAILS

User X makes use of Big Bank's online services regularly. In an Internet Relay Chat (IRC) room one evening he participates in a discussion with an IRC user named John who claims that there was and still may be an error in Big Bank's online transaction services. John was also a user of Big Bank's services until he found the fault.

John explains to user X that in using Big Bank's online services far more regularly than normal users, he came across a bug in the system that, if exploited, would leave Big Bank's clients, their personal details and their money open to attack. John immediately alerted Big Bank

about the bug and trusted they would take care of the issue. Much time passed and the bug issue had still not been addressed by Big Bank.

John decided that in the interest of himself and Big Bank's clients he needed to notify the public of the bug. He set up a Web Site explaining that Big Bank had a bug in their online transaction services and that even though he tried to notify them of the bug, they have done nothing to fix it.

John's statements about Big Bank's bugged online transaction service would create an atmosphere of distrust and uncertainty amongst Big Bank's clients. This would result in a loss of profit. Since Big Bank is a business, in accordance with our definition in section 2, its primary objective is profit. Therefore, to ensure trust amongst Big Banks users, Big Bank immediately distrusts John and any of his statements despite the fact that they are true.

Thousands upon thousands of users (including the people that trust them within their local Web of Trust) will automatically distrust John and his statements about Big Bank.

Depending on who trusts Big Bank (and the extent of this trust), John's statements may go completely unrecognised by anyone at all. Large Semantic Web search engines may trust Big Bank to the extent that they may choose not to index John's Web Site because Big Bank has deemed John to be untrustworthy.

5. CONCLUSION

The aim of this paper is not to discredit businesses and their activities on the Web. We merely point out that businesses are established primarily to make a profit. There have been many cases in the past where businesses have chosen to put the trust of their clients second to that of making a profit, [10, 12, 4, 5] are but a few.

As businesses extend their services onto the Web, and eventually the Semantic Web, most issues of trust will be dependent upon issues of business. Business A will trust business B (and hence the users that trust A will implicitly trust B) only if it is profitable for A to trust B. If trusting B will result in a loss of profit for business A, then once again by definition (since the primary motive of A is to make a profit), A will not trust B.

Khare et al [9] describe a trust based approach amongst parties on the Web as "*essential, since the Web crosses many trust boundaries that old-school computer security cannot even begin to handle*". Khare et al introduce the most fundamental principle behind Trust Management:

"*Be specific*". Users need the ability to define exactly how the trust between themselves and a party is to be configured.

But even if the option was available to users to be so specific in delegating their trust, would this option be explored to its full potential? Khare et al point out that Web surfers are often all too eager to put aside messages the likes of "Show alert each time unsecured data is sent?".

When surfing the Semantic Web or reviewing the report of an agent, will users wish to closely analyse the paths taken to make trust decisions by their software? Perhaps they too will eagerly ignore the message or encourage their agents to ignore the message: "Show alert each time new party is [dis]trusted?'.

The Web of Trust is essential in making the Semantic Web a reality. If users are given, and more importantly make use of, the opportunity to carefully define the way trust is delegated between themselves and their trusted parties, abusing this trust for purposes of profit may not be a possibility.

References

[1] T. Berners-Lee. Information Management: A Proposal. 1989, Available online at http://www.w3.org/History/1989/proposal.html

[2] T. Berners-Lee, J. Hendler and O. Lassila. The Semantic Web. Scientific American, Vol. 284, 5, pp. 34-43, 2001

[3] D. Brickley and S. V. Guha. Resource Description Framework (RDF) Schema Specification 1.0. 2000, Available online at http://www.w3.org/TR/rdf-schema/

[4] T. Burrows. "Telkom Takes EasyInfo To Task." ITWeb, ITWeb, 2002, Available online at http://www.itweb.co.za

[5] T. Burrows. "New EasyInfo Directory Sparks Outrage." , ITWeb, 2002, Available online at http://www.itweb.co.za

[6] G. J. de J Cronje, G. S. du Toit and M. D. C. Motlatla. Introduction to Business Management. Oxford University Press, fifth edition, 2000

[7] D. E. Comer. Computer Networks and Internets. Prentice Hall, second edition, 1999

[8] S. Garfinkel. PGP: Pretty Good Privacy. O'Reilly and Associates, 1994

[9] R. Khare and A. Rifkin. Weaving a Web of Trust. Working Draft, 1997

[10] T. Mainelli. "Hotmail Policy Raises Privacy Concerns." CNN.com, 2002, Available online at http://www.cnn.com

[11] Netcraft. "Netcraft Web Server Survey." 2002, Available online at http://www.netcraft.com/Survey/Reports/

[12] Sapa-AFP. "Bush Backs Probe Into Enron Scandal." Mail and Guardian, 2002, Available online at http://www.mg.co.za

[13] A. Swartz and J. Hendler. The Semantic Web: A Network of Content for the Digital City. Proceedings Second Annual Digital Cities Workshop, Kyoto, Japan, 2001