Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Standards & Interfaces 27 (2005) 257–268

COMPUTER STANDARDS & INTERFACES

www.elsevier.com/locate/csi

Security standards for the semantic web☆

Bhavani Thuraisingham*,[1]

Data and Applications Security, The National Science Foundation, Suite 1125, 4201 Wilson Boulevard, Arlington, VA 22230, United States

Received 6 July 2004; accepted 11 July 2004
Available online 4 August 2004

Abstract

This paper first describes the developments in standards for the semantic web and then describes standards for secure semantic web. In particular XML security, RDF security, and secure information integration and trust on the semantic web are discussed. Some details of our research on access control and dissemination of XML documents are also given. Next privacy issues for the semantic web are discussed. Finally some aspects of secure web services as well as directions for research and standards efforts for secure semantic web are provided.

Keywords: Semantic web; XML; RDF; Ontology; Security; Privacy

1. Introduction

Recent developments in information systems technologies have resulted in computerizing many applications in various business areas. Data has become a critical resource in many organizations, and therefore, efficient access to data, sharing the data, extracting information from the data, and making use of the information has become an urgent need. As a result, there have been many efforts on not only integrating the various data sources scattered across several sites, but extracting information from these databases in the form of patterns and trends has also become important. These data sources may be databases managed by database management systems, or they could be data warehoused in a repository from multiple data sources.

The advent of the World Wide Web (WWW) in the mid 1990s has resulted in even greater demand for managing data, information and knowledge effectively. There is now so much data on the web that managing it with conventional tools is becoming almost impossible. New tools and techniques are needed to effectively manage this data. Therefore, to provide interoperability as well as warehousing between the multiple data sources and systems, and to extract information from the databases and warehouses on the web, various tools are being developed. Consequently the web is evolving into what is now called the semantic web.

☆ The views and conclusions expressed in this paper are those of the author and do not reflect the policies of the National Science Foundation or of the MITRE Corporation.

* Tel.: +1 703 292 8930; fax: +1 703 292 9073.
E-mail address: bthurais@nsf.gov.
[1] On leave from The MITRE Corporation, Bedford, MA, USA.

0920-5489/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.csi.2004.07.002

In Ref. [13], we provided an overview of some directions in data and applications security research. In this paper we focus on one of the topics and that is securing the semantic web. While the current web technologies facilitate the integration of information from a syntactic point of view, there is still a lot to be done to integrate the semantics of various systems and applications. That is, current web technologies depend a lot on the human-in-the-loop for information integration. Tim Berners Lee, the father of WWW, realized the inadequacies of current web technologies and subsequently strived to make the web more intelligent. His goal was to have a web that will essentially alleviate humans from the burden of having to integrate disparate information sources as well as to carry out extensive searches. He then came to the conclusion that one needs machine understandable web pages and the use of ontologies for information integration. This resulted in the notion of the semantic web [4].

A semantic web can be thought of as a web that is highly intelligent and sophisticated and one needs little or no human intervention to carry out tasks such as scheduling appointments, coordinating activities, searching for complex documents as well as integrating disparate databases and information systems. While much progress has been made toward developing such an intelligent web there is still a lot to be done. For example, technologies such as ontology matching, intelligent agents, and markup languages are contributing a lot toward developing the semantic web. Nevertheless one still needs the human to make decisions and take actions.

Recently there have been many developments on the semantic web (see for example, [14,15]). The World Wide Web consortium (W3C) is specifying standards for the semantic web [21]. These standards include specifications for XML, RDF, and Interoperability. However it is also very important that the semantic web be secure. That is, the components that constitute the semantic web have to be secure. In addition, the components have to be integrated securely. The components include XML, RDF and Ontologies. In addition, we need secure information integration. We also need to examine trust issues for the semantic web. It is therefore critical that we need standards for securing the semantic web including specifications for secure XML, secure RDF and secure interoperability.

This paper first provides an overview of the standards for the semantic web. In particular, Tim Berners Lee's description of the various layers that would comprise the semantic web is given. It then focuses on security standards for the semantic web. In particular, XML security, RDF security, and secure information integration will be discussed. We also discuss privacy for the semantic web and then describe secure web services. The organization of this paper is as follows. In Section 2 we provide an overview of the standards for the semantic web. Security standards for the semantic web are discussed in Section 3. We detail our research on XML security further in Section 4. Privacy for the semantic web is addressed in Section 5. Some aspects of secure web services are discussed in Section 6. Finally some direction for further research and standards efforts for secure semantic web are given in Section 7.

2. Standards for the semantic web

Tim Berners Lee has specified various layers for the semantic web (see Fig. 1). At the lowest level one has the protocols for communication including TCP/IP (Transmission Control Protocol/Internet protocol), HTTP (Hypertext Transfer Protocol) and SSL (Secure Socket Layer). The next level is the XML (eXtensible Markup Language) layer that also includes XML schemas. The next level is the RDF (Resource Description Framework) layer. Next come the Ontologies and Interoperability layer. Finally at the highest-level one has the Trust Management layer. Each of the layers is discussed below.

TCP/IP, SSL and HTTP are the protocols for data transmission. They are built on top of more basic communication layers. With these protocols one can transmit the web pages over the Internet. At this level one does not deal with syntax or the semantics of the documents. Then comes the XML and XML Schemas layer. XML is the standard representation language for document exchange. For example, if a document is not marked-up, then each machine may display the document in its own way. This makes document exchange extremely difficult. XML is a markup language that follows certain rules and if all documents are marked up using XML then there is uniform representation and presentation of documents. This is one of the

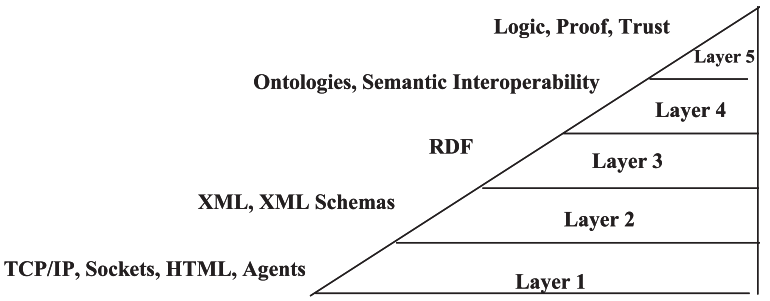

Fig. 1. Layers for the semantic web.

significant developments of the WWW. Without some form of common representation of documents, it is impossible to have any sort of meaningful communication on the web. XML schemas essentially describe the structure of the XML documents. Both XML and XML schemas are the invention of Tim Berners Lee and the W3C (see Ref. [5]).

Now XML focuses only on the syntax of the document. A document could have different interpretations at different sites. This is a major issue for integrating information seamlessly across the web. In order to overcome this significant limitation, W3C started discussions on a language called RDF in the late 1990s. RDF essentially uses XML syntax but has support to express semantics. One needs to use RDF for integrating and exchanging information in a meaningful way on the web. While XML has received widespread acceptance, RDF is only now beginning to get acceptance. So while XML documents are exchanged over protocols such as TCP/IP, HTTP and SSL, RDF documents are built using XML.

Next layer is the Ontologies and Interoperability layer. Now RDF is only a specification language for expressing syntax and semantics. The question is what entities do we need to specify? How can the community accept common definitions? To solve this issue, various communities such as the medical community, financial community, defense community, and even the entertainment community have come up with what are called ontologies. One could use ontologies to describe the various wines of the world or the different types of aircraft used by the United States Air Force. Ontologies can also be used to specify various diseases or financial entities. Once a community has developed ontologies, the community has to publish these ontologies on the web. The idea is that for anyone interested in the ontologies developed by a community to use those ontologies. Now, within a community there could be different factions and each faction could come up with its own ontologies. For example the American Medical Association could come up with its onotologies for diseases while the British Medical Association could come up with its own ontologies. This poses a challenge as the system and in this case the semantic web has to examine the ontologies and decide how to develop some common ontologies. While the goal is for the British and American communities to agree and come up with common ontologies, in the real-world differences do exist. The next question is what do ontologies do for the web? Now, using these ontologies different groups can communicate information. That is, ontologies facilitate information exchange and integration. Ontologies are used by web services so that the web can provide semantic web services to the humans. Ontologies may be specified using RDF syntax.

The final layer is logic, proof and trust. The idea here is how do you trust the information on the web? Obviously it depends on whom it comes from. How do you carry out trust negotiation? That is, interested parties have to communicate with each other and determine how to trust each other and how to trust the information obtained on the web. Closely related to trust issues is security and will be discussed later on. Logic-based approaches and proof theories are being examined for enforcing trust on the semantic web. Note that the layers as evolving as progress is made on the semantic web. For example, more recently a layer in query and rules has been included to support query and rule processing capability. Therefore for more up-to-date information we refer to the work of W3C.

3. Security standards for the semantic web

3.1. Overview

We first provide an overview of security issues for the semantic web and then discuss some details on XML security, RDF security and secure information integration, which are components of the secure semantic web. As more progress is made on investigating these various issues, we hope that appropriate standards would be developed for securing the semantic web. As stated earlier, logic, proof and trust are at the highest layers of the semantic web. That is, how can we trust the information that the web gives us? Closely related to trust is security. However security cannot be considered in isolation. That is, there is no one layer that should focus on security. Security cuts across all layers and this is a challenge. That is, we need security for each of the layers and we must also ensure secure interoperability as illustrated in Fig. 2.

For example, consider the lowest layer. One needs secure TCP/IP, secure sockets, and secure HTTP. There are now security protocols for these various lower layer protocols. One needs end-to-end security. That is, one cannot just have secure TCP/IP built on untrusted communication layers. That is, we need network security. Next layer is XML and XML schemas. One needs secure XML. That is, access must be controlled to various portions of the document for reading, browsing and modifications. There is research on securing XML and XML schemas. The next step is securing RDF. Now with RDF not only do we need secure XML, we also need security for the interpretations and semantics. For example under certain context, portions of the document may be Unclassified while under certain other context the document may be Classified. As an example one could declassify an RDF document, once the war is over. Lot of work has been carried out on security constraints processing for relational databases. One needs to determine whether these results could be applied for the semantic web (see Ref. [9]).

Once XML and RDF have been secured the next step is to examine security for ontologies and interoperation. That is, ontologies may have security levels attached to them. Certain parts of the ontologies could be Secret while certain other parts may be Unclassified. The challenge is how does one use these ontologies for secure information integration? Researchers have done some work on the secure interoperability of databases. We need to revisit this research and then determine what else needs to be done so that the information on the web can be managed, integrated and exchanged securely.

Closely related to security is privacy. That is, certain portions of the document may be private while certain other portions may be public or semi-private. Privacy has received a lot of attention recently partly due to national security concerns. Privacy for the semantic web may be a critical issue, That is, how does one take advantage of the semantic web and still maintain privacy and sometimes anonymity. Note that W3C is actively examining privacy issues and a good starting point is P3P (Platform for Privacy Preferences) standards.

We also need to examine the inference problem for the semantic web. Inference is the process of posing queries and deducing new information. It becomes a problem when the deduced information is something the user is unauthorized to know. With the semantic web, and especially with data mining tools, one can

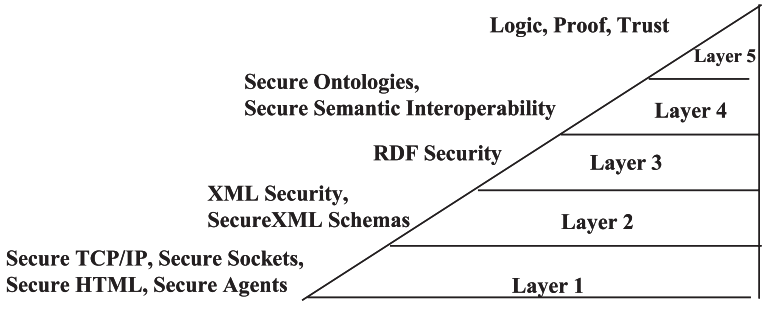

Fig. 2. Layers for the secure semantic web.

make all kinds of inferences. That is the semantic web exacerbates the inference problem (see Ref. [11]). Recently there has been some research on controlling unauthorized inferences on the semantic web. We need to continue with such research (see for example, Ref. [3]).

Security should not be an afterthought. We have often heard that one needs to insert security into the system right from the beginning. Similarly security cannot be an afterthought for the semantic web. However, we cannot also make the system inefficient if we must guarantee one hundred percent security at all times. What is needed is a flexible security policy. During some situations we may need one hundred percent security while during some other situations say 30% security (whatever that means) may be sufficient.

3.2. XML security

Various research efforts have been reported on XML security (see for example, [1]). We briefly discuss some of the key points. XML documents have graph structures. The main challenge is whether to give access to entire XML documents or parts of the documents. Bertino et al. have developed authorization models for XML. They have focused on access control policies as well as on dissemination policies. They also considered push and pull architectures. They specified the policies in XML. The policy specification contains information about which users can access which portions of the documents. In Ref. [1] algorithms for access control as well as computing views of the results are also presented. In addition, architectures for securing XML documents are also discussed. In Ref. [2] the authors go further and describe how XML documents may be published on the web. The idea is for owners to publish documents, subjects to request access to the documents and untrusted publishers to give the subjects the views of the documents they are authorized to see. We discuss XML security in more detail in Section 4.

W3C (World Wide Web Consortium) is also specifying standards for XML security. The XML security project (see Ref. [18]) is focusing on providing the implementation of security standards for XML. The focus is on XML-Signature Syntax and Processing, XML-Encryption Syntax and Processing and XML Key Management. W3C also has a number of working groups including XML Signature working group (see Ref. [19]) and XML encryption working group (see Ref. [20]). While the standards are focusing on what can be implemented in the near-term lot of research is needed on securing XML documents. The work reported in Ref. [1] is a good start.

3.3. RDF security

RDF is the foundations of the semantic web. While XML is limited in providing machine understandable documents, RDF handles this limitation. As a result, RDF provides better support for interoperability as well as searching and cataloging. It also describes contents of documents as well as relationships between various entities in the document. While XML provides syntax and notations, RDF supplements this by providing semantic information in a standardized way.

The basic RDF model has three types: they are resources, properties and statements. Resource is anything described by RDF expressions. It could be a web page or a collection of pages. Property is a specific attribute used to describe a resource. RDF statements are resources together with a named property plus the value of the property. Statement components are subject, predicate and object. So for example, if we have a sentence of the form "John is the creator of xxx", then xxx is the subject or resource, Property or predicate is "Creator" and object or literal is "John". There are RDF diagrams very much like say ER diagrams or object diagrams to represent statements.

There are various aspects specific to RDF syntax and for more details we refer to the various documents on RDF published by W3C. Also, it is very important that the intended interpretation be used for RDF sentences. This is accomplished by RDF schemas. Schema is sort of a dictionary and has interpretations of various terms used in sentences. RDF and XML namespaces resolve conflicts in semantics.

More advanced concepts in RDF include the container model and statements about statements. The container model has three types of container objects and they are Bag, Sequence, and Alternative. A bag is an unordered list of resources or literals. It is used to mean that a property has multiple values but

the order is not important. A sequence is a list of ordered resources. Here the order is important. Alternative is a list of resources that represent alternatives for the value of a property. Various tutorials in RDF describe the syntax of containers in more detail.

RDF also provides support for making statements about other statements. For example, with this facility one can make statements of the form "The statement A is false" where A is the statement "John is the creator of XXX". Again one can use object-like diagrams to represent containers and statements about statements. RDF also has a formal model associated with it. This formal model has a formal grammar. For further information on RDF we refer to the work of W3C reports (see Ref. [6]). As in the case of any language or model, RDF will continue to evolve.

Now to make the semantic web secure, we need to ensure that RDF documents are secure. This would involve securing XML from a syntactic point of view. However with RDF we also need to ensure that security is preserved at the semantic level. The issues include the security implications of the concepts resource, properties and statements. That is, how is access control ensured? How can properties and statements be protected? How can one provide access control at a finer grain of granularity? What are the security properties of the container model? How can bags, lists and alternatives be protected? Can we specify security policies in RDF? How can we resolve semantic inconsistencies for the policies? How can we express security constraints in RDF? What are the security implications of statements about statements? How can we protect RDF schemas? These are difficult questions and we need to start research to provide answers. XML security is just the beginning. Securing RDF is much more challenging.

3.4. Secure information interoperability

Information is everywhere on the web. Information is essentially data that makes sense. The database community has been working on database integration for some decades. They encountered many challenges including interoperability of heterogeneous data sources. They used schemas to integrate the various databases. Schemas are essentially data describing the data in the databases (see Ref. [7]).

Now with the web, one needs to integrate the diverse and disparate data sources. The data may not be in databases. It could be in files both structured and unstructured. Data could be in the form of tables or in the form of text, images, audio and video. One needs to come up with technologies to integrate the diverse information sources on the web. Essentially one needs the semantic web services to integrate the information on the web.

The challenge for security researchers is how does one integrate the information securely? For example, in Refs. [8,10] the schema integration work of Sheth and Larson was extended for security policies. That is, different sites have security policies and these policies have to be integrated to provide a policy for the federated databases system. One needs to examine these issues for the semantic web. Each node on the web may have its own policy. Is it feasible to have a common policy for a community on the web? Do we need a tight integration of the policies or do we focus on dynamic policy integration?

Ontologies are playing a major role in information integration on the web. How can ontologies play a role in secure information integration? How do we provide access control for ontologies? Should ontologies incorporate security policies? Do we have ontologies for specifying the security policies? How can we use some of the ideas discussed in [2] to integrate information securely on the web? That is, what sort of encryption schemes do we need? How do we minimize the trust placed on information integrators on the web? We have posed several questions. We need a research program to address many of these challenges.

3.5. Trust for the semantic web

Recently there has been some work on trust and the semantic web. The challenges include how do you trust the information on the web? How do you trust the sources? How do you negotiate between different parties and develop contracts? How do you incorporate constructs for trust management and negotiation into XML and RDF? What are the semantics for trust management?

Researchers are working on protocols for trust management. Languages for specifying trust management constructs are also being developed. Also there

is research on the foundations of trust management. For example, if A trusts B and B trusts C, then can A trust C? How do you share the data and information on the semantic web and still maintain autonomy. How do you propagate trust? For example, if A trusts B at say 50% of the time and B trusts C 30% of the time, then what value do you assign for A trusting C? How do you incorporate trust into semantic interoperability? What are the quality of service primitives for trust and negotiation? That is, for certain situations one may need 100% trust while for certain other situations 50% trust may suffice.

Another topic that is being investigated is trust propagation and propagating privileges. For example, if you grant privileges to A, what privileges can A transfer to B? How can you compose privileges? Is there an algebra and calculus for the composition of privileges? Much research still needs to be done here. One of the layers of the semantic web is Logic, Proof and Trust. Essentially this layer deals with trust management and negotiation between different agents and examining the foundations and developing logics for trust management.

4. Access control and dissemination of XML documents

Bertino et al were one of the first to examine security for XML (see Refs. [1,2]). They first propose a framework for access control for XML documents and then discuss a technique for ensuring authenticity and completeness of a document for third party publishing. We briefly discuss some of the key issues.

In the access control framework proposed in Ref. [1], security policy is specified depending on user roles and credentials (see Fig. 3). Users must possess the credentials to access XML documents. The credentials depend on their roles. For example, a professor has access to all of the details of students while a secretary only has access to administrative information. XML specifications are used to specify the security policies. Access is granted for an entire XML documents or portions of the document. Under certain conditions, access control may be propagated down the XML tree. For example, if access is granted to the root, it does not necessarily mean access is granted to all the children. One may grant access to the DTDs and not to the document instances. One

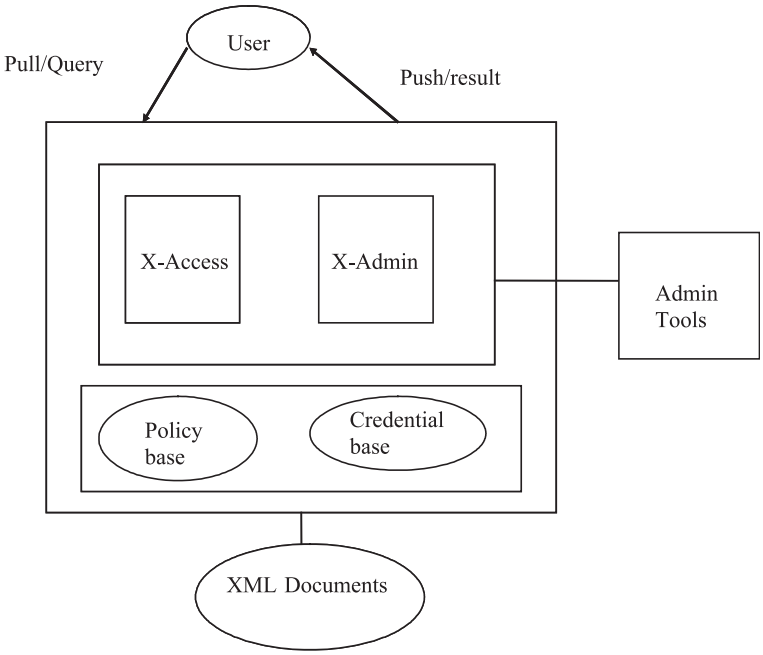

Fig. 3. Access control for XML documents.

may grant access to certain portions of the document. For example, a professor does not have access to the medical information of students while he has access to student grade and academic information. Design of a system for enforcing access control policies are also described in Ref. [1]. Essentially the goal is to use a form of view modification so that the user is authorized to see the XML views as specified by the policies. More research needs to be done on role-based access control for XML and the semantic web.

In Ref. [2] we discuss the secure publication of XML documents (see Fig. 4). The idea is to have untrusted third party publishers. The owner of a document specifies access control polices for the subjects. Subjects get the policies from the owner when they subscribe to a document. The owner sends the documents to the Publisher. When the subject requests a document, the publisher will apply the policies relevant to the subject and give portions of the documents to the subject. Now, since the publisher is untrusted, it may give false information to the subject. Therefore, the owner will encrypt various combinations of documents and policies with his/her private key. Using Merkle signature and the encryption techniques, the subject can verify the authenticity and completeness of the document (see Fig. 4 for secure publishing of XML documents).

As we have stated in Section 3, there are various standards efforts on XML security. The challenge is for the researchers to get their ideas into the standards. There are also a number of commercial products. Therefore, researchers, standards organizations and vendors have to work together to develop appropriate security mechanisms for XML documents. At the same time we need to start research on RDF security and security for the semantic web.

5. Privacy and the semantic web

5.1. Overview

Privacy is about protecting information about individuals. Privacy has been discussed a great deal in the past especially when it relates to protecting medical information about patients. Social scientists as well as technologists have been working on privacy issues. However, privacy has received enormous attention during the past year. This is mainly because of the advent of the web, the semantic web, counter-terrorism and national security. For example in order to extract information about various individuals and perhaps prevent and/or detect potential terrorist attacks data mining tools are being examined. We have heard a lot about national security vs. privacy in the media. This is mainly due to the fact that people are now realizing that to handle terrorism, the government may need to collect data about individuals and mine the data to extract information. Data may be relational or it may be text, video and images. This is causing a major concern with various civil liberties unions (see Ref. [16]).

In this section we discuss privacy threats that arise due to data mining and the semantic web. We also discuss some solutions and provide directions for standards. Section 5.2 will discuss issues on data mining, national security and privacy. Some potential solutions are discussed in Section 5.3.

5.2. Data mining, national security, privacy and the semantic web

With the web and the semantic web, there is now an abundance of data information about individuals

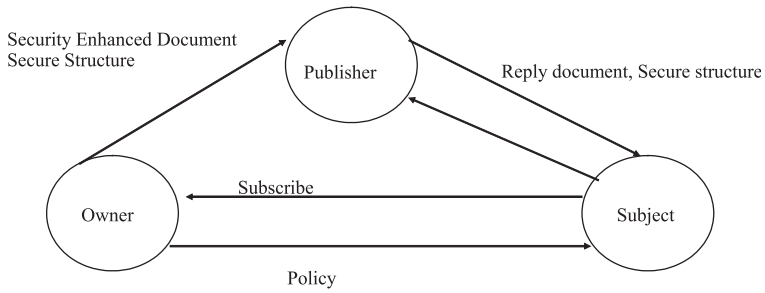

Fig. 4. Secure XML publishing.

that one can obtain within seconds. The data could be structured data or could be multimedia data such as text, images, video and audio. Information could be obtained through mining or just from information retrieval. Data mining is an important tool in making the web more intelligent. That is, data mining may be used to mine the data on the web so that the web can evolve into the semantic web. However this also means that there may be threats to privacy. Therefore, one needs to enforce privacy controls on databases and data mining tools on the semantic web. This is a very difficult problem. In summary, one needs to develop techniques to prevent users from mining and extracting information from data whether they are on the web or on networked servers. Note that data mining is a technology that is critical for say analysts so that they can extract patterns previously unknown. However, we do not want the information to be used in an incorrect manner. For example, based on information about a person, an insurance company could deny insurance or a loan agency could deny loans. In many cases these denials may not be legitimate. Therefore, information providers have to be very careful in what they release. Also, data mining researchers have to ensure that privacy aspects are addressed.

While little work has been reported on privacy issues for the semantic web we are moving in the right direction. As research initiatives are started in this area, we can expect some progress to be made. Note that there are also social and political aspects to consider. That is, technologists, sociologists, policy experts, counter-terrorism experts, and legal experts have to work together to develop appropriate data mining techniques as well as ensure privacy. Privacy policies and standards are also urgently needed. That is, while the technologists develop privacy solutions, we need the policy makers to work with standards organizations so that appropriate privacy standards are developed.

5.3. Solutions to the privacy problem

As we have mentioned, the challenge is to provide solutions to enhance national security but at the same time ensure privacy. There is now research at various laboratories on privacy enhanced/sensitive data mining (e.g., Agrawal at IBM Almaden, Gehrke at Cornell University and Clifton at Purdue University, see for example Refs. [22–24]). The idea here is to continue with mining but at the same time ensure privacy as much as possible. For example, Clifton has proposed the use of the multiparty security policy approach for carrying out privacy sensitive data mining. While there is some progress we still have a long way to go. Some useful references are provided in Ref. [25] (see also Ref. [26]).

We give some more details on an approach we are proposing. Note that one mines the data and extracts patterns and trends. The privacy constraints determine which patterns are private and to what extent. For example, suppose one could extract the names and healthcare records. If we have a privacy constraint that states that names and healthcare records are private then this information is not released to the general public. If the information is semi-private, then it is released to those who have a need to know. Essentially the inference controller approach we have discussed is one solution to achieving some level of privacy. It could be regarded to be a type of privacy sensitive data mining. In our research we have found many challenges to the inference controller approach we have proposed (see Ref. [9]). These challenges will have to be addressed when handling privacy constraints (see also Ref. [17]). Fig. 5 illustrates privacy controllers for the semantic web. As illustrated, there are data mining tools on the web that mine the web databases. The privacy controlled

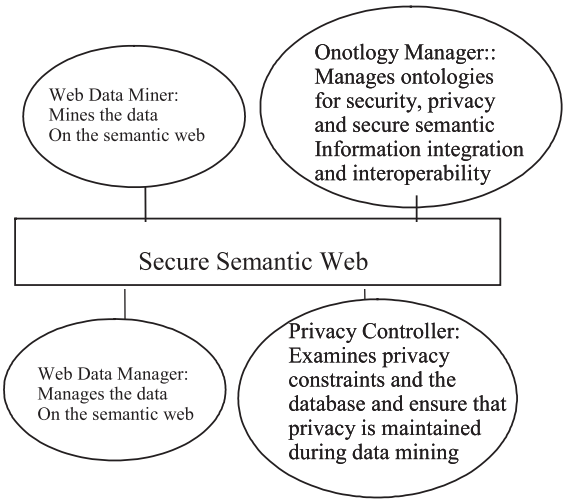

Fig. 5. Privacy controller for the semantic web.

should ensure privacy preserving data mining. Ontologies may be used by the privacy controllers. For example, there may be ontology specification for privacy constructs. Furthermore, XML may be extended to include privacy constraints. RDF may incorporate privacy semantics. We need to carry out more research on the role of onotologies for privacy control.

Much of the work on privacy preserving data mining focuses on relational data. We need to carry out research on privacy preserving semantic web data mining. We need to combine techniques for privacy preserving data mining with techniques for semantic web data mining to obtain solutions for privacy preserving semantic web data mining. In one of our earlier papers in this journal we discussed standards for data mining (see Ref. [12]). We need to develop standards for privacy preserving data mining, which will include standards for data mining as well as standards for privacy.

6. Secure web services

Web services are services such as resource management services, directory services, publishing service, subscription service and various other services that are provided on the web. There has to be some way to describe the communication on the web in a structured and organized way. WSDL (Web Services Description Language) does this by defining an XML grammar for describing network services. As described in Ref. [21], the network services are described as a collection of communication endpoints capable of exchanging messages. WSDL document has various elements including the following: Types, which is a container for data type definition; Message, which is the data being communicated; Operation, which is an action supported by the service; Port Type, which is a subset of operations supported by the endpoints; Binding, which is a concrete protocol and data format specification for a particular port type; Port, which is an endpoint; and Service, which is a collection of endpoints.

So while WSDL is a web services description language what are web services? These are services provided by the web to its users. These could be publishing services, data management services, information management services, directory services, etc. That is, any service that the web provides is a web service and WSDL provides the means to specify the service. Web services is an area that will expand a great deal in the coming years. These services will form the essence of the semantic web.

Now these web services have to be secure. This would mean that we need to ensure that appropriate access control methods are enforced. We also need to ensure that malicious processes do not subvert the web services or cause a denial of service. There has been a lot of work these past few years or so on secure web services. Intrusion detection techniques are being examined to detect and prevent malicious intrusions. Extensions to WSDL for security are proposed. Some details can be found in Ref. [21].

Fig. 6 illustrates a layered architecture that includes secure web services layer for the semantic web. While we show a hierarchy in Fig. 6, it does not mean that all the layers have to be present. For example, secure web services layer may have direct communication with

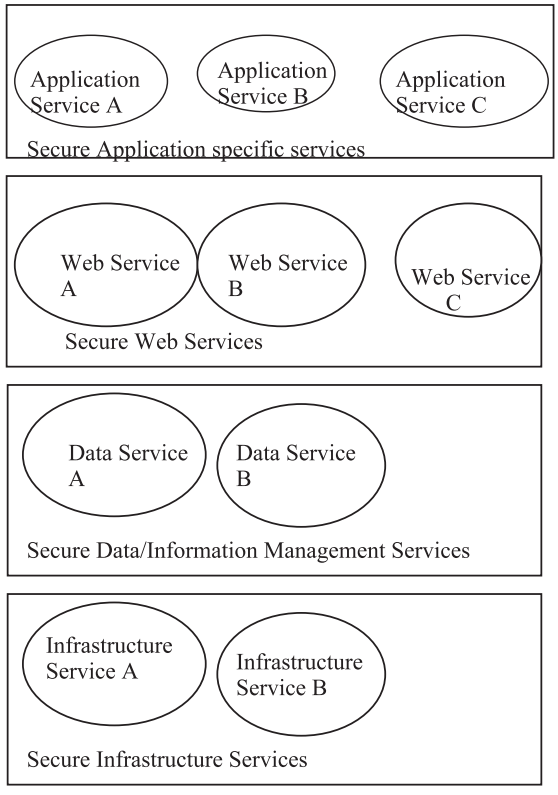

Fig. 6. Layered architecture for secure web services.

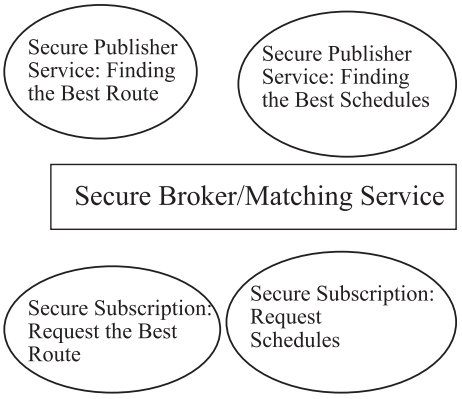

Fig. 7. Example secure web service.

secure infrastructure services layer if there is no need for the services provided by the secure data management services layer. The application services layer provides application specific services such as carrying out banking or filing income tax papers. We also need to ensure end-to-end security. That is, not only each service must be secure, we must also ensure secure interoperability. Fig. 7 illustrates a secure publish and subscribe service that the web provides. This is part of the web services layer and takes advantage of the services provided by the data management layer for getting the necessary data. Standards efforts for web services are already under way with W3C.

7. Summary and directions

This paper has provided an overview of the semantic web and discussed security standards. We first discussed the layered framework of the semantic web proposed by Tim Berners Lee. Next we discussed security issues. We argued that security must cut across all the layers. Furthermore, we need to integrate the information across the layers securely. Next we provided some more details on XML security, RDF security, secure information integration and trust. If the semantic web is to be secure we need all of its components to be secure. We also described some of our research on access control and dissemination of XML documents. Finally we discussed privacy for the semantic web.

There is a lot of research that needs to be done. We need to continue with the research on XML security. We must start examining security for RDF. This is much more difficult as RDF incorporates semantics. We need to examine the work on security constraint processing and context dependent security constraints and see if we can apply some of the ideas for RDF security. Finally we need to examine the role of ontologies for secure information integration. We have to address some hard questions such as how do we integrate security policies on the semantic web? How can we incorporate policies into ontologies? We also cannot forget about privacy and trust on the semantic web. That is, we need to protect the privacy of individuals and at the same time ensure that the individuals have the information they need to carry out their functions. We also need to combine security research with privacy research. Finally we need to formalize the notions of trust and examine ways to negotiate trust on the semantic web. We have a good start and are well on our way to building the semantic web. We cannot forget about security and privacy. Security must be considered at the beginning and not as an afterthought.

Standards play an important role in the development of the semantic web. W3C has been very effective in specifying standards for XML, RDF and the semantic web. We need to continue with the developments and try as much as possible to transfer the research to the standards efforts. We also need to transfer the research and standards to commercial products. The next step for the semantic web standards efforts is to examine security, privacy, quality of service, integrity and other features such as multimedia processing and query services. As we have stressed security and privacy are critical and must be investigated while the standards are being developed.

Acknowledgements

I thank the National Science Foundation and the MITRE Corporation for their support of my research on security issues for the semantic web.

References

[1] E. Bertino, et al., Access Control for XML Documents, Data and Knowledge Engineering, North Holland, 2002, pp. 237–260.

[2] E. Bertino, et al., Secure Third Party Publication of XML Documents, to appear in IEEE Transactions on Knowledge and Data Engineering.
[3] C. Farkas, Inference problem for the semantic web, Proceedings of the IFIP Conference on Data and Applications Security, Colorado, August, 2003.
[4] T. Berners Lee, et al., The semantic web, Scientific American (2001 May) 34–43.
[5] S. St. Laurent, XML, McGraw Hill, New York, NY, 2000.
[6] Resource Description Framework, www.w3c.org.
[7] A. Sheth, J. Larson, Federated database systems, ACM Computing Surveys (1990 September) 183–236.
[8] B. Thuraisingham, Security issues for federated database systems, Computers & Security (1994) 200–212.
[9] B. Thuraisingham, W. Ford, Security constraint processing in a distributed database management system, IEEE Transactions on Knowledge and Data Engineering (1995) 274–293.
[10] B. Thuraisingham, Data Management Systems Evolution and Interoperation, CRC Press, Boca Raton, FL, 1997.
[11] B. Thuraisingham, Data Mining: Technologies, Techniques, Tools and Trends, CRC Press, Boca Raton, FL, 1998.
[12] B. Thuraisingham, C. Clifton, Standards for data mining, Computer Standards and Interfaces Journal (2001) 146–152.
[13] B. Thuraisingham, "Data and applications security" developments and directions, Proceedings IEEE COMPSAC, 2002.
[14] B. Thuraisingham, XML, Databases and the Semantic Web, CRC Press, Florida, 2001.
[15] B. Thuraisingham, The semantic web, in: W. Bainbridge (Ed.), Encyclopedia of Human Computer Interaction, Berkshire Publishers, 2003.
[16] B. Thuraisingham, Web Data Mining: Technologies and Their Applications to Business Intelligence and Counter-Terrorism, CRC Press, Florida, 2003.
[17] B. Thuraisingham, Privacy constraint processing in a privacy enhanced database system, Data and Knowledge Engineering Journal, 2003, accepted for publication.
[18] http://xml.apache.org/security/.
[19] http://www.w3.org/Signature/.
[20] http://www.w3.org/Encryption/2001/.
[21] www.w3c.org.
[22] R. Agrawal, R. Srikant, Privacy-preserving data mining, Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.
[23] J. Gehrke, The research problems in data stream processing and privacy-preserving data mining, Proceedings of the Next Generation Data Mining Workshop, Baltimore, MD, November 2002.
[24] C. Clifton, M. Kantarcioglu, J. Vaidya, Defining privacy for data mining, Proceedings of the Next Generation Data Mining Workshop, Baltimore, MD, November 2002.
[25] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, Privacy preserving mining of association rules, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 2002.
[26] B. Thuraisingham, Data Mining, National Security, Privacy and Civil Liberties, ACM SIGKDD Explorations, December 2002, pp. 1–10.

Dr. Bhavani Thuraisingham is the Program Director for Cyber Trust and Data and Applications Security of the National Science Foundation and has been on IPA to NSF of the MITRE Corporation since October 2001. She is part of a team at NSF setting directions for cyber security and data mining for counter-terrorism. She has been with MITRE since January 1989, where she was the Department Head in Data and Information Management of the Information Technology Division and later the Chief Scientist in Data Management in MITRE's Information Technology Directorate. She has conducted research in secure databases for over 18 years and is the recipient of IEEE Computer Society's 1997 Technical Achievement Award and recently IEEE's 2003 Fellow Award for her work in database security. She is also a 2003 Fellow of the American Association for the Advancement of Science. Dr. Thuraisingham has published over 200 refereed conference papers and over 60 journal articles in secure data management and information technology. She serves (or has served) on editorial boards of journals including IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Dependable and Secure Computing, ACM Transactions of Information and Systems Security, Journal of Computer Security and Computer Standards and Interface Journal. She is the inventor of three patents for MITRE on Database Inference Control and has written six books on data management and data mining for technical managers and is currently writing a textbook on database and application security based on her work the past 18 years. Her research interests are in secure semantic web, sensor information security, and data mining for counter-terrorism.