CS437 XML Webdata Management

Semester Project: Secure XML Publishing

Phase 2 Report

Aaron Powers

Tom Guenther

Matt Nuckolls

Abstract

Secure Publishing of XML data to the web to provide access for multiple users is necessary. Certain users may need access to this data, while other users may not be allowed access. Additionally, publishers need the ability to update documents that have already been published. This paper proposes a method for achieving this, utilizing a publish/subscribe method and xpath querying for partial updates and user access level verification. The first section of this paper introduces the problem statement. Next, the problem is explored further by discussing XML and its security shortcomings. The third section describes the proposed solution to the problem, and work completed thus far to implement the solution. The fourth section describes potential data to user for testing as well as methods to verify that the model is functioning as intended. The final section will present a survey of other research that has been done in the same area.

Problem Statement

Storing data in an XML format is convenient due to the fact that it does not follow a specific schema. The user has the responsibility of defining their own schema which makes data storage in XML very flexible. Data stored in an XML format can be published on the web for distribution purposes. This allows for a convenient method of distributing data to other individuals interested in the data. Publishing XML data on a network is a convenient way to distribute data, however security obviously is a concern. The person publishing may want to allow an individual, or individuals, to access some of their documents or even parts of their documents, while preventing everyone else on the network from accessing their data. A secure method of publishing XML data is needed to prevent access to all portions of a document that an individual is not allowed to access. Additionally, an individual should have the ability to subscribe to xml documents, or document topics. Subscribing to a document would allow an individual to be “pushed” updates as they are published so they do not have to periodically search for new or updated data. Additionally, a publisher may need to update a document that has already been published. A method of providing a partial update is needed to allow a publisher to update a portion of his document without republishing the entire document. Another feature that would be convenient is for newly published partial updates to be automatically pushed to all authorized subscribers to the document.

Analysis of the Problem

XML is a flexible markup language. It is intended to be both machine readable as well as human readable. Though it has a specific structure defined by tags, elements, attributes, etc. it has no specific format nor requirement for tag sets. This flexibility makes it versatile but also vulnerable. Since XML is simply a markup language, it has no intrinsic security mechanisms. Therefore, XML data storage and transmission security must be handled externally to the XML content. XML document encryption is necessary for secure storage and transmission. Furthermore, XML document access must also be handled externally. Again, XML has no intrinsic access controls. Since access and transmission control must be handled externally to the XML document (implicitly: the XML parser), a “3rd party” application may be required to authorize access over a secure transmission protocol such as SSL (secure socket layer).

Proposed Solution

To solve the problem of securely publishing XML data on the web to provide access to authorized users, we propose to implement a simple publisher/subscriber model and an interface by which users can access the system. We intend to store data externally using SQL Lite. Stored data will consist of published XML data, user access level information, and user subscription information. By maintaining this data, newly published or updated documents can be pushed to all authorized subscribed users automatically. When a user accesses the system for the first time, any topic he subscribes to can be stored, as well as all access level information.

To provide a better understanding of the proposed solution, we now describe the layers that it is broken into. The first layer (and the only layer visible to the user) is a simple interface. This can be thought of as a simple web page that allows a user to publish documents, update already published documents, or subscribe to documents. The only purpose of this layer is to pass data and query requests through to the next layer. The second layer of the system is the security, or access control, layer. This layer is responsible for passing published data, user subscriptions, and access levels through to the next layer. When a publisher publishes a document, he also publishes what users or access levels are authorized to access the document. It is the security layer’s responsibility to store this information, while the information that it doesn’t need to store (e.g. the actual data and user subscriptions) is passed through to the next layer. It is also the security layer’s responsibility to verify user access level for results retrieved from the third layer, before returning them to the user. The third and final layer can be seen as a Database Processing Layer. This layer has the responsibility of storing published data into a database as well as storing user subscriptions. This layer also has the responsibility of retrieving stored data and returning it to the security layer. It can be viewed as a layer that handles database input and output. A diagram detailing the model is shown below.

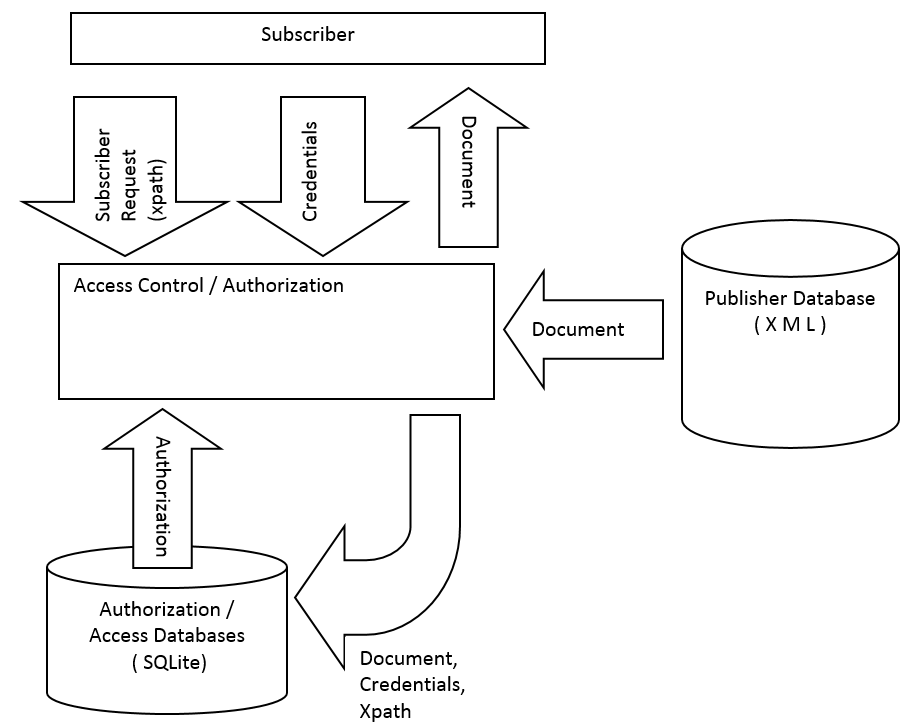
 As described above, the first layer in our proposed solution is an interface to allow a user to publish and subscribe to data. This layer is currently implemented as a simple HTML page in the python language. Running the python program creates the webpage on the local host on port 8025, so simply opening an internet browser and browsing to the local host on that port displays the interface. While not intended for actual use, this simple interface provides a user-friendly, convenient way to test the lower layers of the model which is where the actual functionality takes place.

Figure 1: Secure Publish/Subscribe Model

The responsibility of the second layer of our system is to handle the security aspect of the model. When a publisher publishes data, this layer passes the published data onto the third layer and also stores any users that the publisher provides as an authorized user of the document, or a portion of the document. We have not implemented this aspect yet but plan to require the publisher to provide user names for each document that he has published to indicate which users can access that document. Additionally, the publisher must provide and xpath query. This xpath query defines the specific portions of a document that a user can access. For instance, if a user should be allowed to access multiple parts of a document but not the entire document, the publisher can provide that user’s name along with an xpath query that finds all these parts and nothing else, or that publisher can provide that user’s name with an xpath query more than once (each time with different xpath query) to specify what that user can access. If a user should be allowed access to the entire document, this can also be specified by the publisher. The user names, document ids, and xpath queries are stored in an “authorization database” within this layer. When a user requests a document or topic, the authorization database is scanned to determine what documents and portions of documents the user can access, then this layer requests those documents from the third layer. The authorization layer finally returns the results back to the interface layer. We have also discussed a method of storing the authorization data within each xml published xml document as metadata. This option could potentially simplify the overall model, but the details of the implementation would be more complex to achieve this, so it is unlikely this method will be fully explored at this time.

The third layer in our solution has the responsibility of storing and retrieving the xml data that has been published. As stated earlier this is stored using SQL Lite to keep persistent data even when the program that runs the interface is not running. The XML data is stored along with a topic and a document id to allow partial updates of individual documents as well as searches based on document topic. Partial document updates are possible by utilizing xpath. Similar to providing a way to verify user authorization of partial documents, xpath also provides a way for the publisher to update a portion of a document. In this case, the publisher will need to provide the document id to be updated, the new/modified data, and an xpath query that will return only those portions to be updated from the original document. When this is provided, this third layer retrieves the original document, utilizes the xpath query to find a portion of it, replaces that portion of the document with the new data, and finally replaces the original document in the database with the newly modified document. Of course in the case of updating an existing document, authorization must be checked by the authorization layer to verify that the user is allowed to update the document. It goes without saying that the original publisher can update a document, but it is conceivable that a publisher would want to allow other users to update his document as well. This could be achieved by storing a list of “authorized modifiers” in the authorization database for each document, but this concept has not been explored to date. A diagram of the current proposed partial update method is provided below.

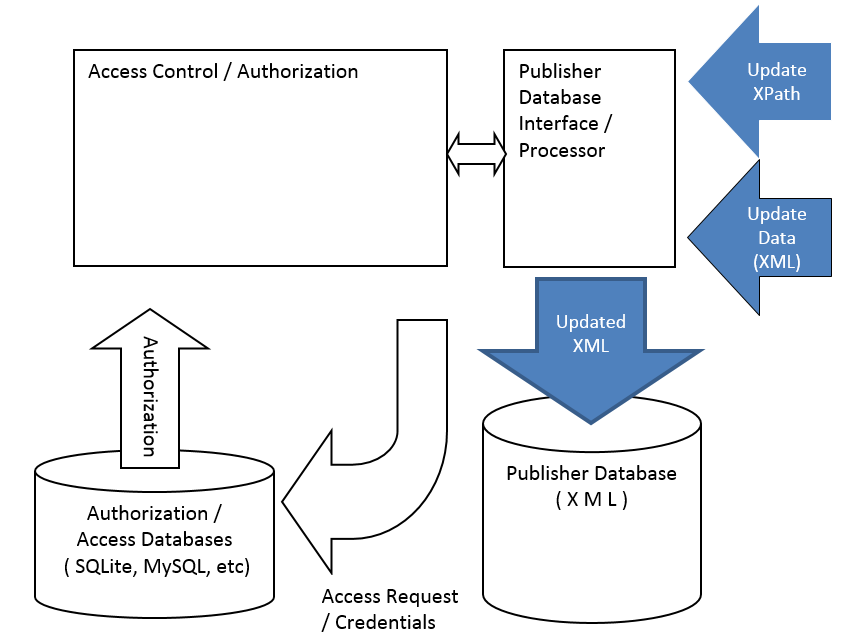


Figure 2: Partial Document Update Method

Test Data and Methodologies

We are currently considering molecular data from a protein database for testing our model. Utilizing this type of data provides a number of benefits. One of which is that there is plenty of real world data, readily accessible on the internet. Another benefit is that this data can provide us with a potential real-world use case scenario to test our model. A real-world scenario where this could be used is sharing protein data between scientists in a lab. Certain scientists may only be allowed to view certain portions of data or certain molecules. This could be based on a number of things such as the scientist’s job title or the level that they are. The molecular data from the protein database comes in a “PDB” file. These files define the contents of the XML document and will be stored in the publisher database in our model.

There are three potential test cases that should prove that the model is working as intended by allowing users to publish and subscribe to data and to do so securely. These three tests cases are publish, subscribe, and update. The publish test case should prove that the publisher database is being populated. The use case is that a user “logs in" to the system and publishes data. The data is then pushed to the publisher database and the XML data is stored as a document in the database. The second test case is to prove that a user that subscribes, gets data that they subscribed to. In this test case, the user logs into the system and subscribes to a topic. The system verifies the user’s access level and then pushes all data that the user is authorized for to the user. Additionally, new data and data changes are automatically pushed periodically to the subscribed user. The third test case is to show that a publisher can do a partial update to a document that has already been published. In this case, the user logs in to the system, and publishes a document that already exists in the database along with an update xpath query. The system first will verify that the user is authorized to make updates to the document in question, then it will retrieve the document and use the provided xpath query to replace the data in the document with the data provided by the publisher. The newly modified document is then stored in the database overwriting the original document.

The interface for running the test cases is as described above. The model will be tested utilized the simple web interface previously described. Users will log in to the web interface to publish and subscribe to documents. Additionally users will provide partial updates via this interface as well. Both publishers and subscribers will need to log into the system so that their name can be recorded. This is utilized to verify authorization for subscribers and to store the publishers name for future document updates. The web page also provides the interface by which users will receive periodic updates for newly published and modified data. At this time, the interface is mainly used to prove that the inner workings of the model are working as intended.

Review of Related Papers

Currently, there is a large amount of work being done in the area of Secure XML Publishing. Mohamed Nabeel and Elisa Bertino published a paper titled “Secure Delta-Publishing of XML Content”. In this paper the authors propose a method of secure XML publishing that supports small incremental changes without the publisher needing to republish the entire paper. Their method also allows that the third party publishers do not need to be trusted due to encryption used by the original publisher.

Erwin Leonardi, Sourav S. Bhowmick, and Mizuho Iwaihara published a paper titled “Efficient Database-Driven Evaluation of Security Clearance for Federated Access Control of Dynamic XML Documents”. The authors of this paper propose a method of Secure XML distribution that utilizes a “policy enforcer” and “data provider”. Essentially, the data provider stores the XML documents as well as the access policy of those xml documents. A user can query the policy enforcer for data from the XML documents which the policy enforcer passes on to the data provider. The data provider runs the query then returns to the policy enforcer the documents found in which the user is allowed to access.

A sophisticated approach to the problem of secure xml publishing was presented in a paper by Mohammad Ashiqur Rahaman, Yves Roudier, and Andreas Schaad titled “A Publish/Subscribe Model for Secure Content Driven XML Dissemination”. This paper presents a very detailed example of a publish/subscribe model and how it could be used in a real world scenario between different organizations. This method ensures that the XML schema as well as the data itself are both confidential, ensures the integrity of transmitted data, and removes the coupling of publishers with subscribers by inducing a dissemination layer between the publishers and subscribers. The method also ensures the XML data cannot be read by the dissemination layer. In this model a publisher first establishes authorization policies that are enforced by the dissemination network then a user sends a subscription request along with his credentials to the dissemination network. When a publisher publishes his encrypted XML document he annotates it with information regarding the “concept” of the document and the dissemination layer forwards the document to the subscribed users. In this model, users only get portions of the document they are authorized for; they can get them automatically by distribution from the dissemination layer; there is confidence in the integrity of the data; and, the data is kept hidden from anyone that is not authorized to see it. This model seems to solve many of the problems surrounding secure XML publishing as well as providing efficiency within the dissemination network layer. However, it does not appear that this model supports incremental updates from publishers.

Another method has been proposed by B. Carminati, E. Ferrari, and E. Bertino in their paper “Secure Third Party Distribution of XML Data”. This is yet another method that provides confidentiality, integrity, and authenticity in secure XML data publishing; but, it also provides completeness. Completeness is provided by ensuring that the users are receiving every portion of the document they are allowed to receive based on their access control policy. Like the method proposed by Rahaman, Roudier, and Schaad, this method does not require that the third party publishers be trusted. However, data still remains confidential. This is achieved by the user receiving keys directly from the data publisher.