Investigation of XML Data Security

Abdelsalam Almarimi[1], Adel Elashheb[2], Faraj Abdulaziz[3], Ezzedin Enbiah[3]

[1] College of Electronic Technology Baniwalid Baniwalid, Libya

[2] College of Civil Aviation & Meteorology, Esbaia, Libya

[3]Higher and Intermediate Institute for Comprehensive Professions, Baniwalid, Libya

Abstract

The extensible Markup Language (XML) is receiving much attention for expressing much of the Web's content and data exchange. However, the use of the internet and multimedia technology these days has encouraged the intruders to copy the contents and illegal use of the data which is available on the internet or any other resources. Therefore, the integrity protection of XML documents is becoming highly important. In this paper, we present a brief overview of XMK related technologies and data security. Moreover, we introduce an experimental example for encrypting XML data based on chaos method.

Keywords: *XML, Data security, Encryption, and Digital watermarking.*

1. Introduction

The recent growth of the Internet and invention of the XML [1] technology data format has created the need for businesses to exchange information, and to interoperate in a uniform way XML has rapidly become the de facto standard for document and data exchange. The extent of its growth is indicated by the growing research in several of the XML and its applications; a series of standards has grown up around it, many of which were defined by the World Wide Web Consortium (W3C). A XML document is a stream of characters, which can be encrypted just like any other document. Any encryption program that can encrypt any document can encrypt an XML file. The problem is that doing so restricts the ability to make use of the full power of XML. Encrypting a XML file will result in a binary stream which is not in the XML format, therefore it needs specialized processing to read and understand. For this purpose, for encrypting an XML file should results into a new XML files [2,3]. The first standard that the W3C produced in the XML cryptography space was XML Signature [4] and the second major XML security specification from the W3C is XML. Watermarking is the process of embedding the information into another object or signal. Its applications mainly used for copyright-protection, through which the owner can prove his ownership or trace any reproduction of the original data. For more effective, the watermark should have the characteristics such as Perceptual Transparency, Robustness, Universality, Capacity, Payload, and Unambiguousness. The rest of the paper is organized as follows. The next section describes XML and related technologies. Section 3 introduces a brief survey of XML Data Security. Section 4 presents XML watermarking Techniques. Section 5 shows an example for experimental results. Finally, conclusions are drawn in Section 6.

2. XML and related technologies

2.1 XML

XML is a subset of the Standard Generalized Markup Language (SGML), an earlier document-structuring language. SGML has been used to create some large information collections such as encyclopedias and multi-volume books, but its complexity has discouraged widespread adoption. SGML is the base technology which specifies a set of rules for putting data structures into a text.
The power of XML as description language lies in the fact that an XML document contains self-describing, hierarchically structured data, and the ability to associate markup terms with data elements. XML appears to subsume HTML and its successor XHTML as the communication language for the Internet. In other words, just as HTML is used to render texts so that they can be processed by humans, XML renders data structures so that they can be processed by computers. Several related standards are greatly increasing the utility of XML for data sharing and management. There are two important concepts of XML document well-formed and validity. An XML document is well formed if it has a root element, every opening tag is followed by a matching closing tag, the elements are properly nested, and any attribute can occur at most once in a given opening tag and its value must be provided. There are many grammar languages that can describe the structure of an XML document. The most common are: DTD (Document Type Definition) and XML Schema. The specification of XML schema is an optional document-structure grammar which is used to make sure the XML document is valid. XML documents can be defined according to a schematic representation by DTD or XML Schema definition. An XML document that conforms to a DTD or XML Schema is called a valid XML document.

2.2 DTD

DTD is a set of rules for structuring an XML document. It is precisely a context-free-grammar for the document. The DTD describes a document type by specifying which tags are allowed, their attributes, and the allowed nesting. Roughly, the DTD corresponds to the schema definition in relational or object-oriented databases. The schema of an XML document may be defined by a DTD. A DTD describes a grammar for semi-structured data description.

2.3 XML Schema

XML Schema was developed to add data types to XML and to provide better document validation than DTDs. XML Schema is also important because DTDs do not support XML Namespaces. Although DTDs have served well for years as the primary mechanism for describing structured information in the SGML and HTML communities, they are considered too limited for many data-interchange applications. For example, DTDs can only specify that elements are text strings. Furthermore, they are not formulated in XML syntax and provide only very limited support for types or name spaces. Therefore, XML Schema was introduced to overcome some of the deficiencies of DTD. XML Schema as data definition language for XML documents has become a recommendation of W3C. An XML Schema Definition (XSD) is an XML-based grammar declaration for XML documents. XML Schema allows very precise definition for both simple and complex data types, and allows deriving new type definitions.

2.4 XML API

XML documents have to be parsed in order to be used by application programs. Therefore, Application Programming Interfaces (APIs) are used to process an XML document by accessing its internal structure. There are three major standardized ways for users to get access to the content of XML documents:

- **Document Object Model (DOM)** is an application program interface (API) for XML instances defined by W3C.
- **The SAX (Simple API for XML)** was the first popular interface for XML programming. A SAX application was a set of event handlers, each called when the parser encountered an element or some text in the document.
- **Java Document Object Model (JDOM)** is a new and open source XML API. JDOM is a tree-based, pure Java API for parsing, creating, and manipulating XML documents. It is a lightweight and fast, and is an XML technology optimized for the Java developer to read, change, and write XML data much more easily than before. JDOM integrates well with both DOM and SAX, and takes the best features from DOM and SAX.

3. XML Data Security

Without any defenses, it is obvious that there are several security threats, and to get around the potential threats, three general security functions: Authentication, Message Integrity and Confidentiality are needed. For this reason we discuss about encryption, signature and key exchange method.

3.1 XML Encryption

XML Encryption is an encryption technology that provides end-to-end security for applications that require secure transmission of XML data. It solves the security problems such as confidentiality, integrity and authentication. XML Encryption [2,3] standards provide mechanisms for the strengths of XML in applications with cryptographic requirements. For the XML Encryption standards provide two complementary capabilities. Firstly, the XML document, much of the markup structure can be modified in ways that do not change the actual meaning of the data. The XML Signature standard provides a means of ensuring changes so that it will not impact on the meaning of a data. Secondly, the standards provide powerful and flexible signature and encryption mechanisms that build on the native abilities of XML. XML Encryption also provides advanced features such as:

- **Partial encryption**: encrypts XML data within specific tags,
- **Multiple encryptions**: encrypts XML data multiple times using different keys,
- **Complex encryption**: encrypt particular portion of the XML tags according to the designation of the recipient.

3.2 XML Signature

XML Signature is an electronic signature technology which is defined to be used in XML data transmission. XML Signature provides a standard for signing an XML document and representing a digital signature in an XML format. XML Signature specification [4] defines electronic signature formats using XML, the creation of electronic signature and rules for the verification process. It solves security problems such as authentication, integrity and non-repudiation. Further to this, XML Signature provides advanced benefits such as partial signature; allows only data contained in specific tags to be signed in the XML document and multiple signature; enables multiple electronic signatures to be included in the XML document.
There are various type of transforms in XML signature. They are simply algorithms that are applied to an XML document and which result in a new XML document. These transforms such as:

- **Canonicalization:** XML Signatures are complemented by Canonical XML, which specifies an algorithm to serialize an XML document so that "equivalent" XML documents produce the same byte sequence [5].
- **Base64:** Base64-encoded content out of an XML document to be transmitted in binary.

3.3 XML Key Managements

The XML Key Management Specification (XKMS) provides an interface between XML applications and a Public Key Infrastructure (PKI) and also it specifies protocols for distributing and registering public keys [6]. This specification eliminates the complex PKI application logic implementation at the client side and allocates trust processing decisions in to separate trust processors. XML Key Information Service Specification (X-KISS) and the XML Key Registration Service Specification (X-KRSS) are two major subparts of the XKMS [7].

4. XML Watermarking

There are many techniques have been proposed for watermarking XML data. R. B. Wolfgang et al. presented two different watermarking schemes on XML data: the selective approach and the compression approach. In the selective approach, the watermarks are randomly distributed throughout the XML document based on a secret key. In this technique, E is a locator candidate and K is the secret key provided by the owner. Using a hash function $H(E,\ \ K)=E\circ K,$ which generates a hash value, h to determine whether E should be marked or not. After determining the marked locators, E_m, watermark the value of E_m according to the data type of E_m. For example, "1000.30000" can be changed to "1000.30001". For textual data, the value of E_m is replaced by a synonym function, denoted as $Synm()$. In the compression approach of Watermarking XML, the basic principle is that the compressed XML data cannot be retrieved directly without the correct decompression [8].

Xuan ZHOU et al [9], they uses the concept of the data usability, which is a measure whether the data can provide useful as well as the correct information to the end user.

F. Frattolillo et al [10,11], presents a watermarking procedure for JPEG images based on the use of protected XML documents. Furthermore, to increase the security and robustness of the procedure, the watermark is repeatedly embedded into an image in the DCT domain at different frequencies and by exploiting both block classification techniques and perceptual analysis. However, the proposed procedure preliminarily estimates the "perceptual capacity" of the coefficients belonging to the luminance DCT blocks by exploiting both block classification techniques and perceptual analysis. In fact, the block classification techniques [12] are applied to select the best DCT coefficients that can be altered without reducing the visual quality. The perceptual analysis is then applied to calculate the "just noticeable difference" values for the DCT coefficients.

Cox, I.J., et al [13], the dual watermark technology based on digital copyright technology is proposed, which, stores the correlation information including the keys and the dual watermarks in an XML document. A new image watermarking technology to spread a digital image with copyright protection is realized successfully on the Internet. In the technique, this embeds two watermarks into one image. The first watermark is the sequence code of products and the second watermark is the meaningful logo. The first watermark is called the proscenium watermark and the second is called the background watermark. The aim of using the proscenium watermark is that it can be detected quickly on the Internet. The aim of using the background watermark is to accurately identify ownership of the product's copyright or to testify to the owner's ID.

5. Experimental work

We consider a case when a user wishing to place an order with an online retailer. The order will include the user's identification, identification of the ordered goods, and payment information (such as a credit card). The user will want to keep private, i.e., encrypt, the payment information. The retailer wishes to authenticate the user to make sure no fraudulent orders are placed, which requires a digital signature. Authenticating the retailer to the user is better handled at the messaging protocol level, and not at the message level. The following XML document is used as original data for XML encryption

```
<?xml version="1.0" encoding="UTF-8" ?>
 <PaymentInfo>
   <Name>John Smith</Name>
   <CreditCard>
     <Issuer>Example Bank</Issuer>
     <Number>4019 2445 0277 5567</Number>
     <Expiration>04/02/2008</Expiration>
   </CreditCard>
 </PaymentInfo>
```

5.1 Encryption of XML documents

The process of XML data encryption consists of the following steps:

1. Read the XML file in the binary form.
2. Select the portion which we want to encrypt.
3. Generate the pseudorandom sequence using chaos which is dependent on the initial values.

$$f(x)=\mu x(1-x)$$

4. Set the threshold $'t'$ for generate the pseudorandom sequence.
5. X-or the selected portion of the XML and binary pseudo-random sequence.

And consider the value of the $\mu = 3.57$, $x_0 = 0.7$ and $t = 5$ to generate the pseudo-random sequence
If credit card number is sensitive information and if the application wishes to keep that information confidential, it can encrypt the CreditCard element:

Encrypted.xml

```
<?xml version="1.0" encoding="UTF-8" ?>
- <PaymentInfo>
  <Oaig>Kolp"Umjvj</Oaig>
- <DrifkvCbtf>
  <Iswwgt>Exeorne!Dcol</Iswwgt>
  <Nuqdgt>405;"4457"138:"8989</Nuqdgt>
  <Extktctjqp>043241212:</Extktctjqp>
  </DrifkvCbtf>
  </PaymentInfo>
```

By encrypting the entire Credit Card element from its start to end tags, the identity of the element itself is hidden

5.2 Decryption of XML document

For decryption of the XML document, also we required the exact value of the initial values and threshold. The process of decryption consists of the following steps:

1. Read the encrypted XML file in the binary form.
2. Generate the pseudorandom sequence using initial condition.
3. Generate the binary pseudo-random sequence using the threshold.
4. X-or the binary pseudorandom sequence and encrypted portion.

Select the for decryption initial value as $\mu = 3.57$, $x_0 = 0.7$ and $t = 5$, then we are able to recover the exact form of the decryption.

If we select $\mu = 3.57$, $x_0 = 0.7001$ and $t = 5$ then the decrypted portion as:

```
<?xml version="1.0" encoding="UTF-8" ?>
  <PaymentInfo>
    <Name>John Smgsf</Name>
    <CreditC_qb />
  </PaymentInfo>
```

The decrypted xml is not same as original if we take different initial value.

6. Conclusions

In this paper, we have briefly discussed the XML data security and various type of XML watermarking techniques. Also, we have introduced an experimental example for XML data encryption based on the chaos technique.
In the future, we plan to investigate the hiding the data in the four level wavelet transform & piecewise linear chaotic map for making it more robust against various types of the attacks.

References

[1] W3C Consortium: Extensible Markup Language (XML). http://www.w3.org/TR/2000/REC-xml

[2] D. Eastlake and J. Reagle. XML Encryption Syntax and Processing. W3C Candidate Recommendation, December 2002. http://www.w3.org/TR/xmlenc-core/

[3] Berin Lautenbach, Introduction to XML Encryption and XML Signature, Information Security Technical Report. Vol. 9, No. 3, 2004, pp. 6-13.

[4] XML-Signature Syntax and Processing. Technical report, W3C, February 2002. http://www.w3.org/TR/xmldsig-core/

[5] World Wide Web Consortium. Canonical XML Version 1.0, Mar. 2001. W3C Recommendation.

[6] Phillip M, Hallam-Baker, and Warwick Ford. XML Key Management Specification (XKMS). VeriSign Inc.

[7] XML Key Management Specification (XKMS). technical report, W3C, March 2001. http://www.w3.org/TR/xkms/

[8] R. B. Wolfgang et al., "Perceptual watermarks for digital images and video," Procs of the IEEE, vol. 87, no. 7, pp. 1108–1126, 1999.

[9] Xuan ZHOU, HweeHwa PANG, Kian-Lee TAN, and Dhruv MANGLA, WmXML: A System for Watermarking XML Data, Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005, pp. 1318-1321.

[10] F. Frattolillo and S. D'Onofrio, "A video watermarking procedure based on XML documents," in Procs of the 13th Int'l Conf. on Image Analysis and Processing, F. Roli and S. Vitulano, Eds., Italy, 2005, vol. 3617 of LNCS, pp. 568–575.

[11] F. Frattolillo and S. D'Onofrio, Exploiting XML Documents to Watermark JPEG Images ," IST 2006 -

International Workshop on ImagingSystems and Techniques Minori, Italy 29 April 2006, pp 34-39.

[12] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in Human Vision, Visual Processing and Digital Display IV, J. P. Allebach and B. E. Rogowitz, Eds., S. Jose, CA, USA, Feb. 1993, vol. 1913 of SPIE Procs, pp. 202–216.

[13] Cox, I.J., Kilian, T., Leighton, J., and Shamoon, T.: 'Secure spread spectrum watermark for multimedia', IEEE Trans. Image Process., 1997, 6, (12), pp. 1673–1687.