A secure platform for content delivery in digital libraries

Mahdi Shamlo[1]*, Nasser Mozayani*[2]*, Homa Tavakoli*[3]
{Shamlo, Mozayani}@iust.ac.ir, Homa_tavakoli@Comp.iust.ac.ir
Iran University of Science and Technology, Tehran, Iran

Abstract

The content delivery system is a major part in DLs (digital libraries). Dealing with different types of digital contents, the content delivery software should be completely compatible with copyright rules and the digital service regulations. Therefore, the content security plays an important role in designing these software systems. We have investigated the security issues both in the content level and in usage level. We put together ensembles of security concepts in a variety of subsystems such as web servers, databases, network connection and user interface. As a result, some more important security aspects and recommendations are presented. We have also implemented a digital content delivery system obtaining a relatively acceptable trade-off point between security and performance factors.

Introduction

In DLs the issue of library resources is revolutionarily changed in different aspects such as publishing, storing, delivering, and using. Accordingly it requires different policies and new methods for providing access to non-physical contents. Nevertheless the digital content is not a virtual object having no service regulations. It has to be regarded as a "*real digital object"* with intellectual properties and copyright regulations. Consequently, the library must be able to control the method and the level of accessing to any digital document.

In a traditional library the access control is guaranteed by limiting replication rules. It is also the intrinsic property of physical objects not being replicated simply

In this study we consider only the digital resources which can be transformed to a non-interactive or print format such as books, journals, reports, maps, pictures, etc. Either these documents are originally published in electronic format or they are transformed from a hardcopy format by scanning, Book-marking, OCR[4], etc. The issue is basically different for other types of resources like multimedia interactive objects, Flash files, etc.

The main objective of this paper is to propose a secure environment for providing access to the former digital contents with maximum security considerations in the framework of library regulations. Here the security includes mainly the confidentiality, the authentication and the authorization. It is worth mentioning that library regulations precisely define the security related to digital contents. Therefore the security here is rather satisfying the copyright rules and respecting e-publishing regulations (Digital Millennium Copyright Act of 1998).

---

[1] BSc student, Computer Engineering Department, IUST
[2] Assistant Professor, Computer Engineering Department, IUST
[3] BSc student, Computer Engineering Department, IUST
[4] Optical Character 'Recognition

Basic viewpoints

There are different levels of threats in a digital content delivery system. In the following, we discussed the fundamental considerations in three levels more in detail.

Platform level

A secured platform is an essential part of each software system. The digital library as a software system should have a secure platform indeed. In client-server architecture and multi-tier systems this security can be viewed in 3 stages.

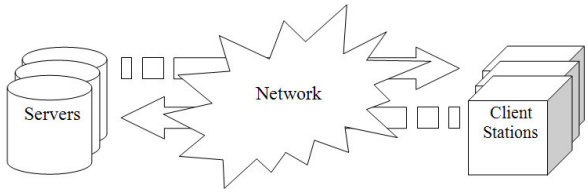

Figure 1 Three stages of platform security

Server stage

This category encompasses threats such as physical damages to server machines, operating system attack via network, penetration to the file system and Trojans (Hartman, Donald, Konstantin, *et al.* 2003).

Fortunately maintaining server security, in this stage, is not a difficult task due to the physical availability of software and hardware components. There are numerous techniques and many hardware-software tools for providing an acceptable level of security. For instance the server-room physical protection, using different firewalls in servers, controlling network ports, and connections and so on are relatively trivial today. Although these methods can not guarantee all aspects of the security but their achievements are practically satisfying.

Access link and network stage

Some well-known threats in this area are packet sniffing, spoofing, session hijacking, denial of service, etc (Curphey M, Scambray J, Olson E, and Howard M. 2003). It is usually very difficult to make sure that the access link is well protected and the communication packets are not received by others on the network. Specially when talking about the Internet which is intrinsically a public network; it is practically impossible to obtain network privacy. Therefore other approaches should be taken; among which the cryptography is a useful solution. Data encryption tools in the network, such as SSL[5] servers are among existing solutions for this problem. VLAN[6] (Curphey M, Scambray J, Olson E, and Howard M. 2003) connection is also a good method for attaining a relatively secured network; whenever possible. Generally, maintaining a secure platform in this stage is more difficult than providing the security in the former stage.

Client stage

Hacking client side application, changing client side standards in order to accessing restricted data, spy-wares, Trojans, etc. are common threats which can be found in client machines .

Maintaining the security in this stage is much more difficult than in previous ones. Skilled hackers can unlock almost any constraint. This is more tangible when talking about web users. Web browsers usually work based on standard HTTP[7] protocol making them very vulnerable and hackers can get through almost any constraints applied to the client side. For example the right click function may not be absolutely disabled in client machine (skilled hackers can unveil any technique used in this regard).

[5] Secure Socket Layer
[6] Virtual Local Area Network
[7] Hyper Text Transfer Protocol

Architecture and logic level

From a logical point of view, a content delivery system may be designed very differently. Data integrity, availability, confidentiality, privacy, and accountability are challenging issues in design process. A trade-off point between the security and cost-performance criterion is not unanimously defined. For example it could be easier to send large blocks of contents to the user's machine and to manage them in client side, but in turn, it would be more difficult to maintain the confidentiality and the privacy of data (Dunn 2003) So depending upon the designed delivery method, different levels of security may be provided.

Implementation level

The topics in this level concern generally the technical issues which are not covered in this paper. Some instances are

- Error handling and proper reacting in non-anticipated cases in the execution of programs.
- Memory insufficiencies and faults may affect the constraints imposed to data delivery system; causing the security holes in the software.
- Unknown components and non-certified programs can bring up faults.
- Code injection, cross site scripting , etc.

Security policy implementation

In the previous section we have globally introduced some major aspects of security in content delivery systems but the main question is in what extent we can implement a secure platform. Is it practically possible to carry out all the aforementioned items in a DL system? What is the tradeoff point between the ideal security and the gained performance such as ease of use, fast access, user friendly GUI[8], maintenance, etc.? It is obvious that more we try to implement security issues; more we get into complex procedures and go toward a complicated system. If we try to maximize the connection security via encryption of verification methods, we will have lower performance in data transfer throughput (Hughes 2006) . The dilemma between having ultra-sophisticated security tools and facilitating the maintenance of systems is also a challenging issue in digital libraries. One cannot afford to dispose very skilful administrators for supporting a relatively simple content delivery system so the latter should be enough simple to be maintained easily by a trained technician (Rathje, McGrory, Pollitt) (Williams and Sears 1998).

We have previously proposed a decoding application for our digital library system. The application was to be installed on the client side and received the encoded data for displaying after a decoding process. Even in such situation some users preferred not to install any plug-ins on their machines. They were, consequently, not able to use our online library system.

As a concluding remark, we can say that security is a relative concept. Depending upon our specific targets and the implementing conditions, we have to define the framework of our secure platform. For instance, the security is differently defined in a university comparing to a military information center. As a second conclusion it is clear that more the security is implemented in server side, better the overall performance obtained in client side (Lim, Goh, Liu 2003).

System parameter	Security control In client side	Security control in network	Security control in server side
Performance	✗	-	✓
Maintainability	✗✗	✗✗	✓✓
Reliability	✓	✓	✓✓
Modifiability	✗✗	✗	✓✓

[8] Graphical User Interface

Runtime efficiency	✗✗	-	✓
Portability	✗✗	✗	✓✓
Ease of use	✗	✗	✓
Flexibility	✗✗	✗	✓
Cost-effectiveness	✗	✗✗	✓✓
Fault tolerance	✗	✓	✓
Robustness	✗	✓	✓✓

Table 1 Some important factors of a secure software system (Chung, Nixon, Yu, et al; Favre 2003).Most of these parameters agree with sever side implementation but in client side security control they can not be achieved generally

Our content delivery solution for digital library systems

As we discussed earlier, DL environments specially *content delivery systems* are of varieties. However the current system is designed to be used in public DLs and universities so the military considerations and special purpose contents are not concerned.

In order to keep up with the security not affecting speed, ease of use and integrity of the system, we should consider various factors to maintain those parameters in developing digital libraries for universal field and universities. Our solution is as follows.

- The quality of delivered contents should be appropriately designed in order to provide enough satisfaction to individual accesses. For example, if a user has a read only access, the system has to provide a tailored service which minimizes the size and the quality of content to fit to the user's need. This will enable us to limit a variety of non-authorized usages.
- *Digital content delivery system* should be based on web application architecture. This will make it possible to use a standard browsing system at the *client side*.
- The delivery content should be in the solid form i.e. Image based contents (Nieder, Steffens, Hemmje 2002; Frey 2000).
- In each query, only the requested part of data has to be extracted from the server and sent to the client. In this way, if the client side is hacked or a failure happened only that small part of data is stolen or lost and the whole record of data is safe on the server.
- We keep the format interchangeable so the system can intervene in case of the generating solid form. For instance a portion of the content could be highlighted or the resolution and size could be modified according to the user's request or based on delivery policies. A stereotype is XML[9].[13] The negative side of this method is the relatively large volume of transferred data which necessarily is in solid form. Despite of mentioned deficiencies this format is a good way due to ease of use and maintaining the security level.
- The activities which deal with the data structure have to be performed on the server side. For example in a full-text search system (Amato, Gennero, Rabili *et al. 2004*), the documents should be selected and processed in the server side and after preparing the appropriate results the separated solid contents are to be sent to the client side.
- Images must be patented through *digital signatures* to prevent misusage. One of the easiest methods is to consider a background image in all solid contents to provide the authenticity of the images. In some cases, watermarking is also a good method to be applied.
- There should be ways to prevent *client side* of misuse and copying without permission of contents. Even though this kind of security is not perfect and still leaves misuse of the system for skilled hackers. There are definitely ways to bypass these limitations and constraints. Some of security controls in the client side includes disabling the *right click*, preventing the cache of the delivered content, preventing the copying of images and viewing the HTML source.

[9] Extensible Markup Language

- The compression technique is also a very effective parameter in gaining user satisfaction. The quality of digital page and the size of transferred documents depend heavily on the compression method. Different compression techniques could be used according to various types of images like texts, pictures, maps, OCRed documents, etc. In the case of user privilege on resolution and quality-due to his bandwidth for example-, the compression can also be modified for attaining these needs (there are varieties of algorithms like DCT[10], vector-based, etc. (Salomon 2000 and Kruck 2004).
- Display frames should be limited in sizes in order to enforce scrolling screen. This will prevent the assembling of scrolled images captured by print-screen utility. Even if possible, one has to devote a huge amount of time to accomplish this task by putting together small parts of image.

A powerful logging system providing a complete set of information about usages such as user id, digital document id, page number, content quality, time and date of access, etc. will help us to prevent misuse through logging tools. To this end, the disposal of proper diagnostic patterns of sequential copying, systematic downloads and unauthorized intrusions are primordial. To visualize this idea, let's imagine a computer agent requests a digital document. The delivered pattern model will provide a result based on frequency, type of browsing, time of request, type of request and other factors of usage; the pattern which help us to recognize the misuse. On top of that we can use advanced data mining methods. We have obtained interesting results using soft computing methods which will published in future reports.

Experimentations

Our experimenting platform is the *Integrated Digital Library System* implemented in IUST[11]. We have added the digital content delivery system –which is called *Online Reading Room*- to this platform. The system is designed to use the existing user account database and to provide them digital contents of more than 8500 electronic document. We have tried to follow the aforementioned security recommendations for delivering a simple, fast and secure system. A sample GUI of the system in client side is illustrated in Figure 2 which comprises the following features.

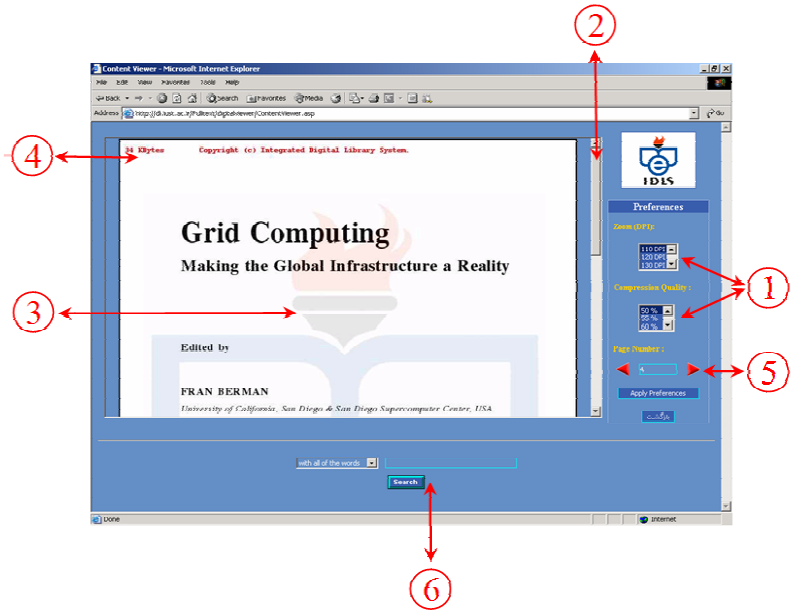

Figure 2

[10] Discrete Cosine Transform
[11] Iran University of Science and Technology

The subsystem for defining the compression quality and zoom level of the displayed pages. Users can determine the required settings according to their connection bandwidth and their preferences.
Content displaying frame. This frame has a fixed size usually smaller than the original target page which makes users to scroll for visualizing the entire page.
Background image showing the logo of hosting library. The image is transparent with a large alpha parameter.
The image size (in kilobytes) of final solid content is dynamically computed and printed here. This allows the user to modify displaying parameters based on his/her needs.
Page request subsystem. User can go to the target page by direct enumerating or by next/previous facility. In this way, the user can never get more than one page at a time.
Search subsystem in server side. Using this utility the user can search in whole document from which only one page is displayed. He/She will visualize another page containing the research results of the same document.

System architecture

The system is designed based on a multi-tier architecture in which the database machine and the storage server machine are not in direct access of users. Queries are carried out by a web server interface machine (Amato, Gennaro, Rabilti *et al.* 2004) (see Figure 3).

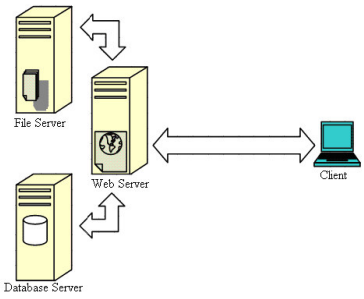

Figure 3 System brief architecture

The hardware platform and execution environment are detailed in appendix.

The experiments are done within two months for about 1300 user with more than 8500 indexed documents and nearly 2 608 225 indexed pages. The usage statistics said that more than %80 of active users logged into the system at least once. Sixty five per cent of the visiting clients had again logged into the system at most six hours after the first login.

The average usage pattern in 24 hours is shown in Figure 4. We have detected some abnormal usages comparing to the average pattern. We have used *Artificial Neural Networks* for extracting meaningful information from these data. The results are very interesting comparing to other case studies (Our results are to be published in future reports).

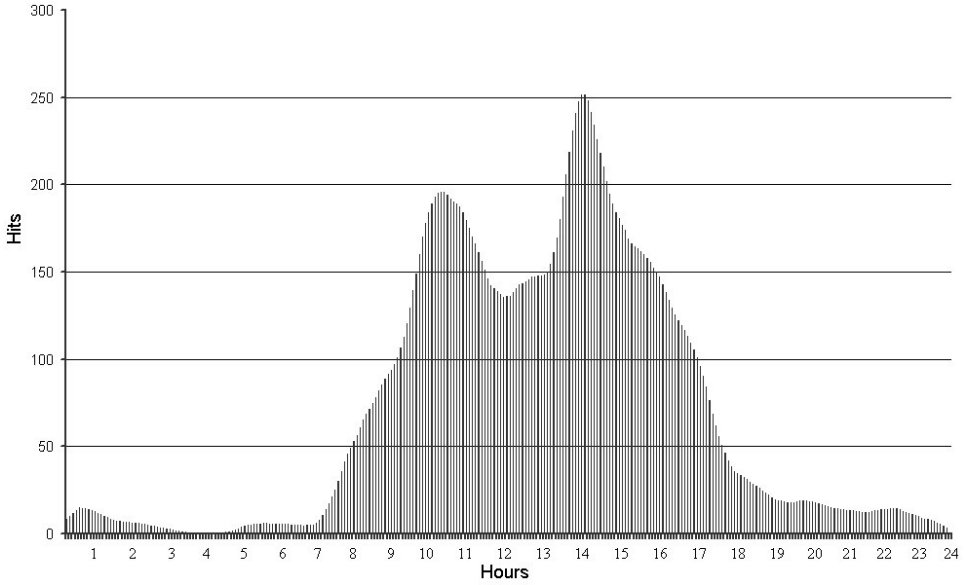

Figure 4 Pattern of contents usage (Daily average)

Conclusion

The concept of security in a d*igital content delivery system* is rather a challenging subject. From an abstract conceptual model to a physically implemented system, there are various levels of security considerations. Depending upon existing conditions and real-world constraints, one has to make some decisions how and in what extent the security should be provided. In this paper we have presented a real implementation which tried to come by a trade-off point between a secure platform and relatively acceptable system parameters including ease of use, maintainability, modifiability, cost-effectiveness, etc.

References

The digital millennium copyright act of 1998, U.S. Copyright Office Summary, December 1998.

Hartman B, Donald J Flinn, Konstantin bezn osov, Shirley kawamoto, 2003, ***Mastering Web Service Security,*** Wily.

Curphey M, Scambray J, Olson E, and Howard M. 2003
Improving web application security threats and countermeasures, Microsoft.

Dunn J, Updated November 2003, **Documentation of The Digital Library Program at Indiana University URLs**, http://www.dlib.indiana.edu/~jenlrile/publications/imageqc/qc.pdf

Hughes L, The Reeves Agency, April 2006: **The value of online books to university libraries**, Elsevier.
By Bente Dahl Rathje, Margaret McGrory, Carol Pollitt, Paivi Voutilainen under the auspices of the IFLA

Libraries for the Blind Section 2005
Designing and Building Integrated Digital Library Systems – Guidelines, The Hague, IFLA Headquarters,
ISBN 90-77897-05-4

Williams, R.D., Sears, B., November 1998
A High-Performance Active Digital Library, *Parallel Computing, Special issue on Metacomputing*.

Lim E P, Goh D, Liu Z, Ng W K, Khoo C, Higgins S E. 2002
"G-Portal: A Map-based Digital Library for Distributed Geospatial and Georeferenced Resources," To appear in proceedings of the *Second ACM+IEEE Joint Conference on Digital Libraries (JCDL 2002)*, Portland, Oregon, USA, July 14-18, 2002

Chung L, Nixon B A, Yu E, Mylopoulos J. 1999
Non-functional requirement in software engineering, Kluwer Academic Publishers, Boston Hardbound, ISBN 0-7923-8666-3 472 pp.
Liliana Favre, 2003
UML and the Unified Process, IRM Press.

C Nieder´ee, U Steffens, and M Hemmje, 2002
Towards Digital Library Mediation for Web Services. In *Proceedings of EurAsia-ICT 2002, First EurAsian Conference of Advances in Information and Communication Technology*.

Frey F, 2000
"Measuring quality of digital masters", Guides to Quality in Visual Resource Imaging, Council on Library and Information Resources, Washington, DC, available at:
www.rlg.org/visguides/visguide4.html

Amato G, Gennaro C, Rabitti F, Milos P S, 2004
A Multimedia Content Management System for Digital Library Applications. *ECDL 2004*

Amato G, Gennaro C, Rabitti F, and Milos P S, 2004
A Multimedia Content Management System for Digital Library Applications. In R. Heery and L. Lyon, editors, *ECDL 2004, Bath, UK, September 13-15, 2004*, volume 3232 of *LNCS*.

Salomon D, 2000
Data compression the complete reference, Springer

Kruk S R 2004
Advanced search and browsing in digital libraries
In *ISWC*

Appendix

Hardware specifications of our test bench are as follows

Web Server:
- CPU: P4 2.7GHz Full Cache
- RAM: 1GBytes
- Storage: 1xHDD 40GByte
- OS: Windows 2000 Server
- Web Server: MS IIS
- Web Application language: COM+, ASP

File Server:
- CPU: P4 1.7GHz
- RAM: 256MByte
- Storage: 4xHDD SCSI 1.2TBytes – Raid 5
- OS: Windows 2000 Server
- File system: NTFS

Storage Server:
- CPU: Dual Xeon 2.7GHz Full Cache
- RAM: 1GBytes
- Storage: 2xHDD SCSI 1.2TBytes - Raid 1
- OS: Windows 2000 Server
- DBMS: MS SQL Server 2000 SP3