8

Security and Digital Libraries

Edward Fox and Noha ElSherbiny
Virginia Tech
USA

1. Introduction

Security is an important issue in digital library design. Security weaknesses in digital libraries, coupled with attacks or other types of failures, can lead to confidential information being inappropriately accessed, or loss of integrity of the data stored. These in turn can have a damaging effect on the trust of publishers or other content providers, can cause embarrassment or even economic loss to digital library owners, and can even lead to pain and suffering or other serious problems if urgently needed information is unavailable (Tyrväinen, 2005).

There are many security requirements to consider because of the variety of different actors working with a digital library. Each of these actors has different security needs (Chowdhury & Chowdhury, 2003). Thus, a digital library content provider might be concerned with protecting intellectual property rights and the terms of use of content, while a digital library user might be concerned with reliable access to content stored in the digital library. Requirements based on these needs sometimes are in conflict, which can make the security architecture of a digital library even more complex.

The design of the security architecture of a digital library must go beyond simply adding one or a few modules to a previously designed system. This is because there may be security holes in pre-existing modules, and because difficulties can arise when attempting to integrate the modules. The security architecture of a digital library must be designed so that security concerns are handled holistically. A security system designer must view the whole architecture and consider all of the applicable security factors when designing a secure digital library. The nature of a security attack may differ according to the architecture of the digital library; a distributed digital library has more security weaknesses than a centralized digital library.

Security attacks can be categorized as physical attacks and logical attacks (Stallings, 2006). A physical attack involves hardware security where keys, locks, cards, and visitor monitoring is used. A logical attack involves an attack on the content or digital library system. We focus on the logical attacks and software security of digital libraries.

2. Security issues with digital libraries

According to the DELOS Reference Model (Candela et al., 2007) there are 6 main concepts in a digital library universe: content, user, functionality, architecture, quality, and policy. Each of these concepts has security issues that affect it.

2.1 Content

The content of a digital library includes the information objects that a digital library provides to the users. Some of the security issues involved are integrity and access control. Integrity requires that each object/resource has not been altered or changed by an unauthorized person. Access control encompasses two security requirements. The first is authentication where the user must log into the system while the second is confidentiality, which means that the content of an object is inaccessible by a person unless they have authorization. Not all digital libraries are free; often content is provided to digital library users for a certain fee, whereupon access control is needed to protect the content. Further, some content is inappropriate for some users, or targeted to particular user groups; there are a whole host of such other reasons for access control.

Logical attacks such as hacking and message tampering can affect the integrity and confidentiality of the content. Improved information access in digital libraries has raised many issues that affect the management of digital libraries. Content Management, or more specifically Digital Rights Management, refers to the protection of content from the different logical security attacks and issues relating to intellectual property rights and authenticity.

2.1.1 Digital rights management

DRM provides content protection by encrypting the content and associating it with a digital license (Tyrväinen, 2005). The license identifies the user allowed to view the content, lists the content of the product, and states the rights the user has to the resource in a computer readable format using a digital rights expression language (DREL) or extensible Rights Markup Language (XrML) that also describes constraints and conditions.

There are 7 technologies used to provide DRM (Fetscherin & Schmid, 2003). Table 1 summarizes the DRM components and supporting technology.

Each of these components involves mechanisms used to provide DRM:

- **Encryption**: Encryption techniques such as symmetric and asymmetric ciphers can be used to provide access control; public-key encryption is used in payment systems that control how and by whom the content is used.

  Symmetric ciphers using DES, 3DES, AES, and RC4 algorithms require the use of a shared secret key to encrypt data before it is sent. At the receiver's end the cipher text is decrypted using the same secret key. Symmetric ciphers depend on both the sender and receiver knowing the shared key.

  Asymmetric ciphers use a pair of keys, public and private, for each of the sender and the receiver. The public keys of both the sender and the receiver are known but the private key is kept secret. If encryption is performed using the public key then only the private key can be used for decryption and vice versa.
- **Passwords**: Stored strings must be matched by users desiring access.
- **Watermarking**: Characters or images are added to reflect ownership. Steganography is used to conceal data inside audio, video, or images (Johnson & Jajodia, 1998). Different watermarking techniques have different aims; some watermarks might be visible while others invisible. Some watermarks are reversible (Mintzer et al., 1997); it depends on the desired use of the watermark and what is being protected.
- **Digital signature**: Asymmetric encryption can be used. Likewise, hash algorithms such as MD5 and SHA can be used to create a signature (Stallings, 2006).

Component	Protection Technology
Access and usage control	Encryption (e.g., symmetric, asymmetric), passwords
Protection of authenticity and integrity	Watermarks, digital signatures, digital fingerprints
Identification by metadata	Allows description of an object in suitable categories, covering the digital content, rights owner, and conditions.
Specific hardware and software	Includes all hardware and software used by the end-device through which the digital content is being played, viewed, or printed.
Copy detection systems	Search engines, which search the network for illegal copies and use watermarking.
Payment systems	Can be seen as a certain type of protection technology as it requires user registration, or credit card authentication, which also require a trust relationship between the content provider and the customer.
Integrated e-commerce systems	DRMS must include systems which support contract negotiation, accounting information, and usage rules.

Table 1. DRM Components and Protection Technologies, adapted (Fetscherin & Schmid, 2003)

- **Digital fingerprint**: Digital fingerprints are a more powerful technique involving digital signatures and watermarking. The creator of the content creates a unique copy of the content marked for each user; the marks are user-specific hence called fingerprints. Should a user illegally distribute the content, the creator can use search robots to find those copies (Schonberg & Kirovski, 2004).
- **Copy detection systems:** Search engines also can help locate such copied objects. Copy-detecting browsers can protect digital content too.
- **Payment systems:** Users must divulge personal information to pay for content. Installing payment systems can help protect digital content.

There is no standard mechanism for providing DRM, mainly due to the lack of regulations (Chowdhury & Chowdhury, 2003), however there are various systems and protocols introduced to provide content management and support fair usage policies.
There is a tradeoff between security and performance. Nadeem and Javed use a Pentium-4, 2.4 GHz machine running Microsoft Windows XP operating system, encrypt 20527 bytes to 2323398 bytes of data using DES, 3DES, and AES. For 20527 bytes of data it took 2 seconds to encrypt using the DES algorithm and 4 seconds to encrypt using the AES algorithm (Nadeem & Javed, 2005). It can be seen that the more complex the encryption algorithm the longer it takes to encrypt the data. In another study, encrypting data with the RSA algorithm using a key size of 1024 took 0.08 milliseconds/operation on an Intel Core 2 1.83 GHz processor under Windows Vista in 32-bit mode, while using a key size of 2048 took 0.16 milliseconds/operation (Dai, 2009).

2.2 User

The User in a digital library refers to "the various actors (whether human or machine) entitled to interact with digital libraries" (Candela et al., 2007). Digital libraries connect the different actors with the information they have and allow the users to consume old or generate new information. Security issues relating to the users of a digital library intersect with content issues discussed above. A main logical security issue relating to users and content is access control. Different access control requirements arise for distributed systems (Tolone et al., 2005) to ensure both confidentiality and authentication:

- Access control must be applied and enforced at a distributed platform level, so should be scalable and available at various levels of granularity.
- Access control models should allow a varied definition of access rights depending on different information and must be dynamic where changes to policies are easily made and easy to manage.
- "Access control models must allow high-level specification of access rights."(Tolone et al., 2005)

Digital library users may need to be authenticated before they can access content. Global/universal identification may not suffice. A service provider that provides content based on a non-identity based criteria like age will not benefit from global identification because there is no way to verify the authenticated user's personal information. Usernames and passwords are not efficient ways to provide authentication.
One of the most widely used authentication protocols is Kerberos. It (Neuman & Ts'o, 1994) is a client–server model, which secures communication with servers on a local network. Developed at MIT in the 1980s to provide security across a large campus network, it is based on the Needham-Schroeder protocol and has now been standardized and included in many operating systems such as UNIX, Linux, Windows 2000, NT, XP, etc.
Kerberos is used as an authentication protocol in cases where attackers monitor network traffic to intercept passwords. It secures communication, provides single sign on and mutual authentication, and does not send a user's password in the clear on an insecure network.
An alternative solution suitable for digital libraries (Winslett et al., 1997), is to represent information about an individual using credentials. Credentials are "abstract objects which contain statements expressing knowledge or information from a definite context." Credentials do not specify direct information about a client and their attributes, they

describe the local environment and context in which the requests originate (Ching et al., 1996).
Digital credentials can be used as a means of authentication in providing DL access control (Winslett et al., 1997). Two agents can be used to assist in the management: a personal security assistant and a server security assistant, to manage digital credentials using a client/server model. The server must notify the client of the credentials required for the current request. The client then sends its credentials for authentication. The client must have some trust of the server to give its credentials, which raises privacy issues.
The personal security assistant is used to obtain credentials on behalf of the client, store the credentials, parse and interpret the required credentials, and manage the acceptance policies (Winslett et al., 1997). A server security assistant is available to specify the credential acceptance policies and their usage.
There is a tradeoff between flexibility and security that must be considered when choosing an access control model, as is discussed below.

2.2.1 Access matrix model

This conceptual model specifies the rights that each subject possesses for each object (Tolone et al., 2005). Actions on objects are allowed or denied based on the access rights specified. There are 2 implementations of the AMM:

- An Access Control List provides a direct mapping of each object the subjects are allowed to access, and their usage rights (owner, read, or write).
- A Capability List defines the objects each subject is allowed to access and the usage rights.

Access control lists and capability lists are not suitable for distributed systems. Their limitations lead to multiple problems (Nagaraj, 2001). ACL provides limited expressibility of policies. Any change in the policies will propagate in the system/application. Authentication in a system that uses ACL solely is a problem because using username & password in a distributed system is not practical. In a distributed system, administration of the system should be decentralized by delegation to reduce the overhead. The owner of the object specifies a policy in ACL. If an overall policy is specified by an entity higher than the object owner, then conflicts may occur in the access rights. The number of administrative entities in a distributed system can be very large. Not all the administrators may have trust amongst themselves, resulting in incorrectly defined policies. For example, admin A may trust B but not C, however B may trust C. If A were to define policy for B then it would be implicitly applicable to C, causing problems.

2.2.2 Role-based access control

Role-based access control involves policies that regulate information access based on the activities the users perform. Such policies require the definition of roles in the system: "a set of actions and responsibilities associated with a particular working activity" (Sandhu & Samarati, 1994). Permissions are assigned to roles instead of individual users. Specifying user authorization involves 2 steps: first assigning the user to a role, second defining the access control that the role has over certain objects.
RBAC is easier to manage and is more extensible than ACL. However RBAC doesn't flexibly handle constraints, where a user with a specific role may need specific permission on an

object. An example of RBAC architecture addressing key limitations is OASIS (Bacon et al., 2003), for use in distributed systems. Role management in OASIS is decentralized and service specific. OASIS is integrated with an event-based middleware that notifies applications of any environmental changes. Roles are parameterized by applications and services to define their client roles, and to enforce policies for role activation and service invocation within each session. Role membership certificates (RMC) are returned to each user on successful login, to be used as a credential to activate other roles (Bacon et al., 2003).
RBAC is suitable for use with digital libraries because it supports decentralized architectures and varying roles, however RBAC doesn't allow for the definition of different roles in a collaborative group.

2.2.3 Task based access control

The Task based access control model extends subject/object access control by allowing the definition of domains by task-based contextual information (Tolone et al., 2005). Steps required to perform the task are used to define access control; the steps are associated with a protection state containing a set of permissions for each state, which change according to the task. TBAC uses dynamic management of permissions.
TBAC systems are limited to defining contexts in relation to activities, tasks, or workflow progress. Since it is implemented by recording usage and validity of permissions, therefore, TBAC requires a central access control module to manage permissions activation and deactivation in a just-in-time fashion.

2.2.4 Team based access control

RBAC doesn't address cases where group members of different roles want to collaborate in a single group. The TMAC model defines collaboration by user context and object context. "User context provides a way of identifying specific users playing a role on a team at any given moment" (Tolone et al., 2005) while object context defines the objects required.
TMAC offers the advantages of RBAC along with ability to specify fine-grained control on users and on object instances. A scalable access control data structure can be used with large collections, applying concepts of team based access control, focusing mainly on the access control data structure, and employing an access control framework called Document Access Control Method (DACM) with a Document Storage System (DocSS) (Gladney, 1997). DACM allows the decentralized administration of privileges, the definition of different rule sets to control a single collection, and different delegation patterns as models.
Current object access control policies use an array of rules to record the privileges each subject is allowed to each object. This is impractical to manage in the large data collections found in digital libraries. DACM solves this problem by finding symmetries in a permission function to allow a brief expression without losing important distinctions.

2.2.5 Content based access control

Another approach to access control models involves defining models according to content. This approach is applicable in digital libraries and distributed systems (Adam et al., 2002), where the access rights to the user are dynamic and may change with each login. Content

based access control policies are very well suited for digital libraries and distributed systems. Recent research has proposed different models; most use digital credentials for authentication, but vary in the definition/storage of the policy.
An important content based access control model (Ferrari et al., 2002), introduces a content-based access control system, Digital Library Authorization System, that utilizes the Digital Library Authorization Model (DLAM). Subject, object, and privilege sets can't be used to define policies in digital libraries mainly because DLs are dynamic with large collections of data and subjects. It defines access control policies based on subject qualifications and characteristics. DLAM provides a means to specify the qualifications and characteristics of subjects. It uses content dependant and independent access control and allows the definition of policies with varied granularity.

2.3 Functionality

The concept of functionality encompasses the services that a Digital Library offers to its users (Gonçalves et al., 2008). The minimum functions of a Digital Library include adding new objects to the library or searching and browsing the library and other functions relating to DL management. A security attack that can affect the functionality of the Digital Library is a Denial of Service attack, which can affect the performance of the system and prevent users from accessing the system.

2.4 Architecture

Digital libraries are complex forms of information systems, interoperable across different libraries and so require an architectural framework mapping content and functionality onto software and hardware components (Candela et al., 2007). There are various models for architecture, e.g., client-server, peer-to-peer, and distributed. All these require the protection of the communication channels between 2 parties, where sensitive data might be transferred (Kohl et al., 1998). Securing the connections involves different layers - Internet, transport, or application layer - depending on the architecture of the system.
The distributed model is scalable and flexible. It is useful when building a digital library with changing content from different sources and offers potential for increased reliability. The security requirements for a distributed digital library are challenging, since the content and operations are decentralized. Fault tolerance and error recovery are issues that affect a distributed system. Replication is used to increase the availability of the system. While this approach solves problems with denial of service attacks, it complicates the protection of the content because a replica of the content exists.
The client-server model doesn't have the same security problems as a general distributed model, however, it presents a major security weakness, the server being a single point of failure. Attacks are concentrated on one server rather than on the multiple replicas of a distributed model.

2.5 Quality

The content and behavior of a Digital Library is characterized and evaluated by quality parameters. Quality is (Gonçalves et al., 2007) a concept not only used to classify functionality and content, but also used with objects and services. Some of the parameters

are automatically measured and are objective while others are considered subjective; some are measured through user evaluations.

2.6 Policy

Policy is the concept that represents the different regulations and conditions that govern the interaction between the Digital Library and users. Policy supports both extrinsic and intrinsic interactions (Candela et al., 2007) and their definition and modification.
Examples of security issues relating to policies include providing digital rights management, privacy, and confidentiality of the content and users, defining user behavior, and collection delivery.

3. Summary

Digital libraries should be secure. This is an important quality that affects all aspects, as has been shown above using the DL characterization of the DELOS Reference Model (Candela et al., 2007). We also can summarize and elaborate upon this point using another framework for DLs (Goncalves et al., 2004).
The 5S framework supports Societies and their needs, covering all aspects mentioned above about Users and related Policies, as well as Quality (Gonçalves et al., 2007). Since Societies cover software actors, agents, components, modules, etc., this also encompasses related Architectural issues. Thus, security with regard to Societies covers issues like client/server, commerce, identity, peer-to-peer, privacy, rights, roles, teams, and trust.
Scenarios cover functions, operations, requirements, services, and tasks. Examples (Gonçalves et al., 2008) include access, access control, authentication, browsing, copying, denial of service attacks, encryption, payment, recovery, searching, usage, and watermarking.
Spaces cover distributed aspects, as well as representations related to 1D, 2D, 3D, and higher dimensional spaces. These include feature, measure, metric, probability, vector, and topological spaces – used throughout computer and human systems.
Structures cover all types of organization, including data structures and databases, with lists (e.g., access control or capability), graphs, and networks. Structures are overlaid on other constructs in the 5S framework, especially on Streams. Thus, documents are structured streams, while protocols involve scenarios applied to structured communication streams. Structures and Streams cover all types of content, and the many security issues related, including digital rights management, fingerprints, and watermarks.
Clearly, DL security support can be complicated, but the above discussion should help readers organize their thinking and make sure that DL systems meet security requirements.

4. References

Adam, N. R., Atluri, V., Bertino, E. & Ferrari, E. (2002). "A Content-Based Authorization Model for Digital Libraries." *IEEE Transactions on Knowledge and Data Engineering* 14(2) : 296-315.

Athanasopoulos, G., Fox, E., Ioannidis, Y., Kakaletris, G., Manola, N., Meghini, C., Rauber, A. & Soergel, D. (2010). A Functionality Perspective on Digital Library

Interoperability. *Research and Advanced Technology for Digital Libraries, Proc. 14th European Conference, ECDL2010, Sept. 6-10*. Glasgow: 405-408

Bacon, J., Moody, K. & Yao, W. (2003). "Access control and trust in the use of widely distributed services." *Software Practice & Experience (Middleware)* 33(4): 375 - 394.

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V. & Schuldt, H. (2007). The DELOS Digital Library Reference Model. http://www.delos.info/index.php?option=com_content&task=view&id=345

Ching, N., Jones, V. & Winslett, M. (1996). Authorization in the Digital Library: Secure Access to Services across Enterprise Boundaries. *Third International Forum on Research and Technology Advances in Digital Librarie*. Washington, DC, IEEE: 110 - 119

Chowdhury, G. & Chowdhury, S. (2003). *Introduction to Digital Libraries*, Facet Publishing.

Dai, W. (2009). "Speed Comparison of Popular Crypto Algorithms." Accessed in 2010, from http://www.cryptopp.com/benchmarks.html.

Ferrari, E., Adam, N. R., Atluri, V., Bertino, E. & Capuozzo, U. (2002). "An Authorization System for Digital Libraries." *The VLDB Journal* 11(1): 58 - 67.

Fetscherin, M. & Schmid, M. (2003). Comparing the Usage of Digital Rights Management Systems in the music, film, and print industry. *Proceedings of the 5th International Conference on Electronic Commerce*. Pittsburgh, Pennsylvania, ACM

Gladney, H. M. (1997). "Access Control for Large Collections." *ACM Transactions on Information Systems (TOIS)* 15(2): 154 - 194.

Gonçalves, M. A., Fox, E. A., & Watson, L. T. (2008). "Towards a Digital Library Theory: A Formal Digital Library Ontology." *Int. J. Digital Libraries* 8(2): 91-114

Goncalves, M., Fox, E., Watson, L. & Kipp, N. (2004). "Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries." *ACM Transactions on Information Systems (TOIS)* 22(2): 270 - 312.

Gonçalves, M. A., Moreira, B. L., Fox, E. A., & Watson, L. T. (2007). ""What is a good digital library?" - A quality model for digital libraries." *Information Processing and Management* 43(5): 1416-1437

Johnson, N. F. & Jajodia, S. (1998). " Exploring Steganography: Seeing the Unseen." *IEEE Computer* 31(2): 26-34.

Kohl, U., Lotspiech, J. & Nusser, S. (1998). Security for the Digital Library - Protecting Documents Rather than Channels. *Ninth Workshop on Database and Expert Systems*. Vienna, Austria: 316 - 321

Mintzer, F., Lotspiech, J. & Morimoto, N. (1997) "Safeguarding Digital Library Contents and Users." *D-Lib Magazine* 3(7/8), July/August 1997, http://www.dlib.org/dlib/december97/ibm/12lotspiech.html.

Nadeem, A. & Javed, M. Y. (2005). A Performance Comparison of Data Encryption Algorithms. *1st International Conference on Information and Communication Technologies*. Karachi, Pakistan, IEEE: 84 - 89

Nagaraj, S. V. (2001). Access Control in Distributed Object Systems: Problems with Access Control Lists. *Tenth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*. Cambridge, MA, IEEE: 163 - 164

Neuman, C. & Ts'o, T. (1994). Kerberos: An Authentication Service for Computer Networks. *IEEE Communications Magazine*, IEEE. 32: 33 - 38

Sandhu, R. S. & Samarati, P. (1994). Access Control: Principle and Practice. *IEEE Communications Magazine*, IEEE. 32: 40 - 48

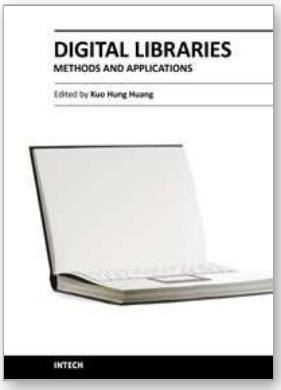
Schonberg, D. & Kirovski, D. (2004). Fingerprinting and Forensic Analysis of Mutimedia. *Media Management*. New York, USA, ACM

Stallings, W. (2006). *Cryptography and Network Security*, Pearson Prentice Hall.

Tolone, W., Ahn, G.-J., Pai, T. & Hong, S.-P. (2005). "Access Control in Collaborative Systems." *ACM Computing Surveys* 37(1): 29 - 41.

Tyrväinen, P. (2005) "Concepts and a Design for Fair Use and Privacy in DRM." *D-Lib Magazine* 11(2), February 2005, http://www.dlib.org/dlib/february05/tyrvainen/02tyrvainen.html.

Winslett, M., Ching, N., Jones, V. & Slepchin, I. (1997). Assuring Security and Privacy for Digital Library Transactions on the Web: Client and Server Security Policies. *IEEE International Forum on Research and Technology Advances in Digital Libraries (ADL)*. Washington, DC, IEEE: 140 - 151



Digital Libraries - Methods and Applications

Edited by Dr. Kuo Hung Huang

ISBN 978-953-307-203-6
Hard cover, 220 pages
Publisher InTech
Published online 04, April, 2011
Published in print edition April, 2011

Digital library is commonly seen as a type of information retrieval system which stores and accesses digital content remotely via computer networks. However, the vision of digital libraries is not limited to technology or management, but user experience. This book is an attempt to share the practical experiences of solutions to the operation of digital libraries. To indicate interdisciplinary routes towards successful applications, the chapters in this book explore the implication of digital libraries from the perspectives of design, operation, and promotion. Without common agreement on a broadly accepted model of digital libraries, authors from diverse fields seek to develop theories and empirical investigations that to advance our understanding of digital libraries.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821