

Security of Heterogeneous Content in Cloud Based Library Information Systems Using an Ontology Based Approach

Mihai DOINEA^{1,2}, Paul POCATILU¹

¹Department of Economic Informatics and Cybernetics,
Bucharest University of Economic Studies, Romania

²Department of Information Technology and Mediatheque
The Romanian Academy Library
mihai.doinea@ie.ase.ro, ppaul@ase.ro

As in any domain that involves the use of software, the library information systems take advantages of cloud computing. The paper highlights the main aspect of cloud based systems, describing some public solutions provided by the most important players on the market. Topics related to content security in cloud based services are tackled in order to emphasize the requirements that must be met by these types of systems. A cloud based implementation of an Information Library System is presented and some adjacent tools that are used together with it to provide digital content and metadata links are described. In a cloud based Information Library System security is approached by means of ontologies. Aspects such as content security in terms of digital rights are presented and a methodology for security optimization is proposed.
Keywords: Cloud Computing, Virtualization, Library Information System, Ontology

1 Introduction

Cloud computing, alongside with mobile technologies, represents an important trend in the development of today's applications. Of this matter, cloud computing is also a research topic in all academic and research institutions in the field of computer science and information technology, example of such projects are presented in [2], [3] and [17].

The paper is organized as follows.

The first section, *Generalities in Cloud-based Systems*, presents the main concepts related to cloud computing. The section also presents the characteristics of public, private and hybrid clouds and provides an example of a private cloud implementation.

The section *Content Security Concerns in Cloud Applications* focuses on security issues raised when using cloud-based applications.

In chapter *Cloud Based Information Library System Implementation* is presented a type of ILS developed by Exlibris along with the adjacent tools that such system is interacting with. An example of an ILS architecture is presented and various tools used along to manage digital content.

The chapter five called *Security Approach Using Ontologies in ILS* is suggesting a security approach using ontologies in order to

improve the content protection feature of these types of systems. Ontologies are used in order to classify content and assign different levels of vulnerability. Based upon these levels, DRM controls implement a restriction policy specific to an ILS system.

The paper ends with conclusions and future work.

2 Generalities in Cloud Based Systems

Cloud computing consists of three service models, as presented in [1], [2] and [3]:

- Infrastructure as a service (IaaS);
- Platform as a service (PaaS);
- Software as a service (SaaS).

IaaS is related to virtualization and provides services for access to hardware (physical or emulated) in terms of computing, storage and networking.

PaaS is oriented to application development and deployment software engines, libraries, compilers etc.). The cloud vendor provides the required environments for clients' applications.

Through *SaaS*, clients have access to applications provided by the cloud vendor. The applications could be for general use (like a word processor or a spreadsheet application) or a dedicated service (like an ERP). Usually,

the client uses a Web browser to access the applications.

Another important characteristic of cloud computing is represented by the deployment models [4]. In a *public cloud*, services and resources are available to everyone and several clients share the infrastructure. In a *private cloud*, the clients don't share the resources with anyone. There is also the *hybrid cloud*, a mix between public and private cloud.

From a technical perspective, several key concepts related to cloud computing have to be presented.

The cloud is based on *nodes* that are physical or virtual machines (VMs). The nodes are grouped in *clusters*. The virtualization is controlled by a *hypervisor* that creates and monitors virtual machines running on a physical server. As examples of hypervisors are mentioned Hyper-V, KVM, VirtualBox, VMWare, Xen etc.

The clients of cloud applications can run on mobile devices, on desktop computers or on servers. This also leads to a new domain, Mobile Cloud Computing as highlighted in [5] and represents an important development field.

2.1 Public Cloud Computing Solutions

One of the most important advantage of using cloud computing is the organizations don't have to invest in datacenters, infrastructure and software, they will be only the services they use [2], [6].

Today, there are numerous cloud solutions vendors. According to [7], in Q3 2014 Amazon is the biggest cloud provider (27%), followed by Microsoft (10%), IBM (7%), Google, salesforce, and Rackspace. Cloud vendors are specialized on a specific service or they sell several services. Table 1 presents several services of the most important cloud vendors of this specific market.

Table 1 Examples of cloud providers and their services

Provider	IaaS	PaaS	SaaS
Amazon [8]	Amzon EC2 Amazon S3	Amazon Elastic Beanstalk	Available on AWS Marketplace
Google [9]	Google Compute	Google App Engine	Google Apps
HP	HP Public Cloud	HP Helion Public Cloud Application Platform as a Service	Available on HP Software Experience Center
IBM	IBM Cloud Managed Services SoftLayer	IBM Bluemix	IBM Solution Provider - IBM Software as a Service
Microsoft [10], [11]	Windows Azure	Windows Azure	Office 365
Oracle [12]	Oracle Compute	Oracle Cloud PaaS	Oracle Applications Cloud Oracle Analytics Cloud Oracle ERP Cloud etc.
Rackspace	Managed Infrastructure		
Salesforce.com		Salesforce platforms	Sales force automation and CRM

The prices vary per service and per unit of resource (hours, storage, traffic etc.). Most of public cloud vendors offer trials or a free of charge period for trying their services.

2.2 Private Cloud Computing Solutions

For an independent developer there are several open source solutions for implementing a private cloud as Google

Ganeti, Eucalyptus, OpenStack, CloudStack etc.

Google Ganeti is used for managing clusters of virtual servers [13]. Ganeti depends on virtualization platforms such as Zen and KVM and it includes functions for:

- Storage management;
- Installation of operating systems;
- Virtual systems control;
- Migration between clusters.

Eucalyptus [13] is open-source software used to build private clouds. The clouds are compatible with Amazon's AWS API. Eucalyptus comprises of several levels: nodes, cluster, cloud and user interface (UI) and application programming interface (API).

OpenStack is another option for building a

private or public cloud. Current release is Juno and it consists of several modules such as [14]:

- OpenStack Compute (Nova);
- OpenStack Dashboard (Horizon);
- OpenStack Identity (Keystone);
- OpenStack Networking (Neutron);
- OpenStack Block Storage (Cinder);
- OpenStack Orchestration (Heat);
- OpenStack Data Processing (Sahara).

OpenStack can be tested on a computer with VM by using the dedicated distribution devstack. Figure 1 presents the Horizon dashboard of an OpenStack installation. VirtualBox is the environment used to run this implementation of OpenStack.

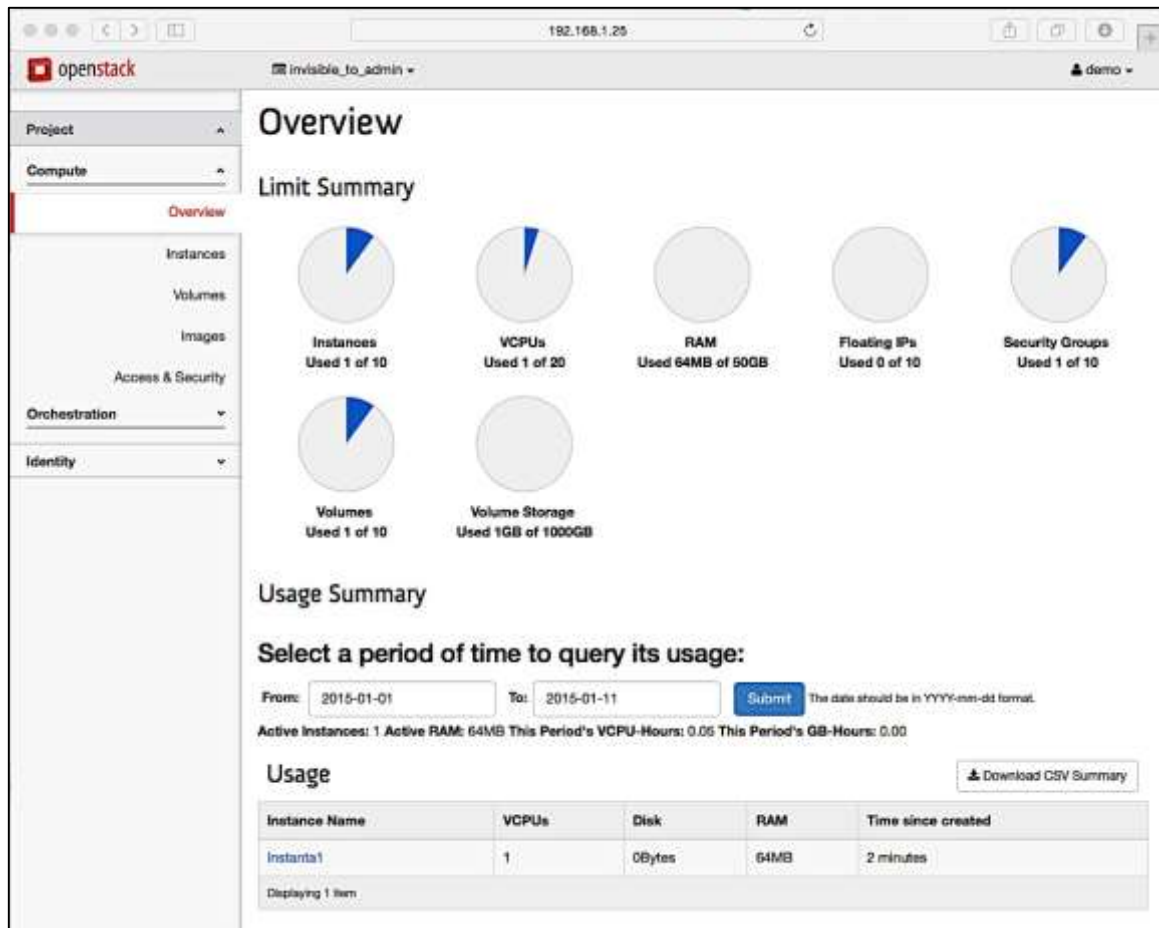


Fig 1. OpenStack dashboard

OpenStack provides several services like compute, image and storage [15].

An information library system can be implemented on a private cloud solution based on its current infrastructure and presented

solutions represent real alternatives. Another option is to use a public cloud services, depending on the existing requirements.

3 Content Security Concerns in Cloud Applications

Security represents an important aspect when using cloud-based applications. Security is an important area of research and technological development for cloud computing, and is supported by European Union funded projects [18].

Cloud based systems represent also an important part of the context of BigData and security issues do not avoid tangling things in this particular area. As stated in [21], one of the major security issues of BigData is represented by the risk associated with the intrusiveness with regards to the personal profile of a user. It is very simple in the current context, to profile a person from their digital records without their consent.

Security measures are split between the organization and the cloud provider. Their weight depends on the type of deployment: private, public or hybrid cloud.

For a private cloud managed by the organization, all the required security measures to be taken by the IT department, including network security. In a public cloud, the cloud provider applies most of the security measures. Also, cloud services users have an important role on selecting passwords and using their credentials (sharing with other colleagues, writing down on post-its etc.).

Regarding the security concerns, these are related to the user and to the cloud provider also, Figure 2.

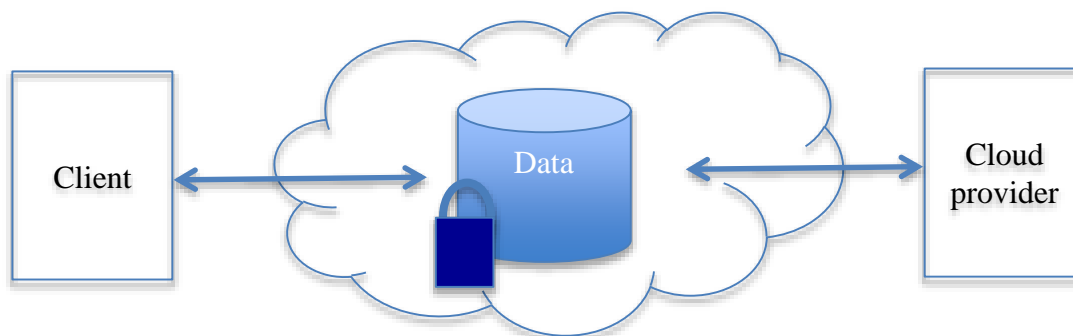


Fig 2. Data security concerns

From the *user* point of view the main security issues could be:

- data and services availability; there could be period of services unavailability due to unexpected caused (power failure, untested applied updates etc.);
- unwanted access to data; unauthorized users could access and change sensitive data stored in cloud;
- data integrity; stored data can be corrupted due a power, communication for storage access failure.

On the users' side, they have to manage the credentials carefully, to choose a strong password and to change it regularly, to announce the support team regarding any unusual behavior or changes etc.

Regarding the *cloud provider*, the main security concern are related to:

- data access; an unauthorized user (from

outside or inside the cloud organization) could access client data or users' connection data;

- services and data availability; internal or external causes could lead to unavailability of the cloud provider's services, that leading to unhappy clients:

Content security is assured by using encryption, monitoring access, implementing network security requirements, making data backups, spreading data on different clusters, etc.

Services runs on virtual machines similar to sandboxes in order to assure that there are no unwanted interaction between the clients of these services [19].

Every cloud solution vendor provides several security mechanisms in their software and infrastructure. For example, AWS instances, especially when used in Virtual Private Cloud

(VPC), can be associated to security groups [20].

Regarding the client applications, these vary from lightweight applications like Web browsers (for mobile or desktop) to dedicated client applications, developed for a specific service. Also, the protocols used for data transfer influences the security model of the application. Proprietary protocols requires specific security measures, having in mind that they are used in a special context and where other protocols are not usable.

4 Cloud Based Information Library System Implementation

An Information Library System, also known as an ILS, must provide multiple service solutions for different types of activities underrun in libraries, museums or archives, based on the support of different software tools that are interconnected with each other. Such powerful tools for ILS systems are developed by Exlibris with an experience of more than 20 years starting with mainframe software dedicated to library services, continuing with distributed systems implemented in client-server architecture and now developing a cloud based infrastructure for library management. LIBISnet is a library network of more than 30 members such as government institutions, public and private organizations. As a service supplier such network must be involved in a lot of projects to be able to identify, design and implement reliable and complete solutions for library systems. Projects such VEP, Europeana inside, Flandrica or Open Vlacc which is one of LIBIS network's biggest ongoing project contribute to the community with ideas and knowledge in order to find as many efficient solutions as possible for the ILS. Open Vlacc represents a bibliographic pool organized as a centralized catalog. Combined with local data and holdings it creates the so called PBS, Provincial Library System, used for public libraries. Open Vlacc is a central catalog of public libraries within the Flanders region being automatically feed by CDR and Boekenbank records.

In the PBS network there are around 30

members of the Belgian "Vlaams-Brabant" province. The PBS network supports access to the online collections, users can see the location of the digital material and submit a loan request. The PBS network is based on an ILS, Aleph v21 running as a client – server architecture, which manages the entire network, structuring the content into different databases according to each one's functionality. The bibliographic records are stored in PBS01 database with records structured in different logical sets based on their physical location.

An example of a cloud based Information Library System is ALMA developed by Exlibris which aims to be the successor of the current client-server product called Aleph which handles a large number of documents, around 7.3 million with a number of copies around 2.3 million. The PBS network is aiming for ALMA integration in order to offer support via a SaaS, System as a Service, platform in the near future.

LIBIS as a development partner of ALMA started the integration of this Unified Resource Management system early in 2014, having on top a Unified Resource Discovery and Delivery system, Primo based, called Limo, as depicted in the Figure 3.



Fig. 3. Architecture of the Library Systems

ALMA is designed to support all library processes starting with data selection, acquisition, metadata management, fulfillment, digitization, resource sharing and ending with external integration into other types of systems.

The main advantages of the ALMA implementation are represented by the shifting from a client-server architecture to the cloud based paradigm, on which the most important focus is given to the electronic resources aiming new expectations regarding the availability characteristic and a more powerful statistic engine which could reveal the areas that need a strategic approach in order to gain efficiency. An important feature of ALMA is considered to be the user driven acquisition process based on which libraries will tend to buy and include in their catalogs, titles that are requested by the users and less materials that are not relevant for their target group. This can be achieved also based on another important feature, the analytics engine, which, due to its multi-tenancy characteristic, can easily output relevant statistics about what materials are the most viewed and benchmarking about system performances.

A cloud based library information system is based upon other services that can provide rich content as input for the system. Besides the content that can come as a result of a digitization process, the ILS needs also the related metadata in order to be able to describe the electronically ingested material. OCR, Optical Character Recognition, and NER, Name Entity Recognition, are tools created inside the SUCCEED project (Support Action Centre of Competence in Digitization) for digitization purposes. The scope is to improve the OCR tools for historical texts in an automated manner as possible. The NER tools was designed for persons, organizations and locations identification inside the text that was processed with OCR tools. The workflow implemented in order to achieve the proposed results has the following stages:

- Digitization – the stage in which the physical materials are digitized, transformed from their physical form into electronic materials stored as images;
- Attestation – create a ground truth for the OCR evaluation; the ground truth represents how a page must look;
- Set evaluation – create training and test sets using the attestation output;

- Quality improvement – the stage that trains the OCR system to recognize special characters by using a special dictionary provided by the INL, Institute of Dutch Lexicology; the output of this stage represent the model used in OCR process;
- Executing OCR – the actual process of extracting characters from an image document.

When implementing the workflow the results were around 80% correct recognition and also around 80% name entities found which demonstrates the success of this project. There are also some drawbacks of the system on which more attention must be given, in terms of book customization. Some stages need to be reevaluated when changing the digital material in order to recalibrate the system to the new book format.

The digitization process starts with tools like OCR and NER and the digital material is stored in archives or presented to the end users. For creating metadata in ALMA in order to connect digital content with associated records, the specialists make use of Digiorder using the Filemaker application. The application is used to describe the content of a book such as details of figures and tables, to number automatically the pages from the book, to add certain notes that appear in the original book and can't be easily reflected in the digital material. The application also describes the chapters' structure creating an automatically content that serves as guidance for mapping the digital content reflected by the scanned images. The output is ingested in LIAS which is an archiving instrument based on Rosetta, the DigiTool successor for online visualization.

The term LIAS refers to the LIBIS application/services stack for archives. The central digital repository for archiving is integrated with domain-specific metadata management system for providing archiving capabilities for libraries and as well for museums. LIAS implements a hierarchical structure, provides content-specific delivery mechanisms and enforces access rights policies to prevent unauthorized access to the

digital content. LIRIAS is an academic archiving tool for researchers' publications affiliated to academia or research institutions. LIRIAS is an open access tool ranked worldwide in the ranking of the institutional repositories. This tool is used to archive all types of research output to which it attributes a unique identifier called a handle, worldwide visible. Also for digital preservation purposes, Rosetta tool is used to enable the university to meet long-term preservation needs of its digital content from libraries, archives and museums.

In order to facilitate content delivery to the other cultural projects such as Europeana inside project, a so called Metadata Interoperability Framework, MIF, can be used. The feeding process involves metadata definition, preview and validation feature, data push services based on Sword as well as OAI-PMH based data pull services and also mapping and transformation support. LIBIS also developed the LibisCoDe that supports transformation services from MARC to EDM and LIDO to EDM to be able to facilitate a successful metadata ingesting process of records exported from their system into the Europeana. This features are provided as services for CMS integration or they can be used from a REST client.

LIMO is another example of a successful implementation of Exlibris Primo product. It helps users to search for printed and electronic publications from different sources. So the Limo implementation unifies search across all of the library resources, even across external resources from other content management systems. Using Limo one can search through the LIBISnet catalogue, the full academic repository called Lirias and the Primo Central containing data harvested from different publishers. As a discovery system Limo makes use of a simple user interface which doesn't affect the user search query. It uses the Apache Lucene Core, a high-performance, full-featured text search engine defining a custom XML format for rules normalization called PNX, Primo XML file. Limo defines facets objects through which content can be filtered by different criteria that

comprehensively describes a set of data.

5 Security Approach Using Ontologies in ILS

The primary role of an ILS is to manage all the electronic and non-electronic resources within a cultural institution in terms of the following undertaken processes acquisition, storage, retrieval, processing and sharing.

An ILS is based on data stored in systems especially designed to allow the most important operations upon them. ILS information can be used for analysis, retrieval, classification, usage, dissemination, extraction, knowledge generation.

The information from an ILS can be accessed through an **Online Public Access Catalog**, called OPAC which allows users to retrieve, filter, save or export the results provided by the online interface.

An important feature of an Information Library System is the ability to provide input, in terms of information, to specialized software equipment, such as data mining systems in order to create knowledge. A virtual library is a nursery for obtaining knowledge, so valuable in the actual context, Figure 4.

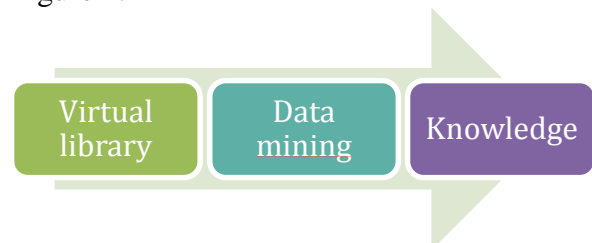


Fig. 4. Knowledge generation flow

The security is very important in this type of systems because any mismatch between the data stored in the virtual library and the actual reality can lead to propagations that could alter important processes that are based on those data.

So the following security characteristics must be maintained for such systems:

- integrity – data can't be altered by an external entity that has not the permission to do so;
- availability – the feature that allows users to dispose of information at any time from any place, if the systems allows it;

- confidentiality – data that is inaccessible to users that have not the right to see it, is protected by passwords or is stored in an encryption form;
- authenticity – says either the data is authentic or not, if the source from which was retrieved is actually the true owner of it;
- non-repudiation – the ability to create an indissoluble relation between the owner and the data that's shared;
- possession – is the feature that allows an owner to be in control of its data at any moment in time;
- utility – assures that data can be used if the owner is able to access it, regardless of other restrictions that are applied to it.

The use on ontologies comes to optimize a security aspect that is very sensitive to the external perception of the ownership of the information retrieved from an ILS. That means to create an instrument that can track digital materials from its source to whatever location it was used in. This will compel others that make use of the retrieved information to include references to the true owner if they want to include it in their work. Ontologies are complex instruments that serve specific purposes which include a lexical approach.

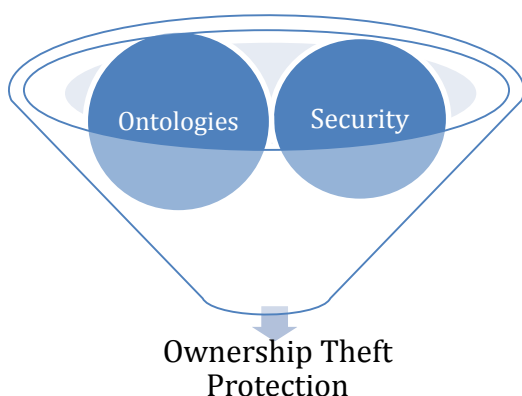


Fig. 5. Ontology usage in ILS security

The most important functions of an ontology are:

- describe a set of concepts;
- illustrate the relations on which the above concepts are based upon;

The ontologies in this security approach is used for the following purposes:

- identifying the correct meaning of a concept in a given context based on the stored concepts and the relations between them;
- determines the major domains that information constituting the virtual library are referring to.

The methodology that describes the implementation of ontology in order to improve a vital security aspect of an ILS consists of the following elements:

- the steps according to which the ontology is used to optimize security;
- the library information system along with all its data;
- security controls for digital rights management.

The algorithm steps for applying an ontology on a library information system in order to determine the major domains of importance are presented in Figure 6.

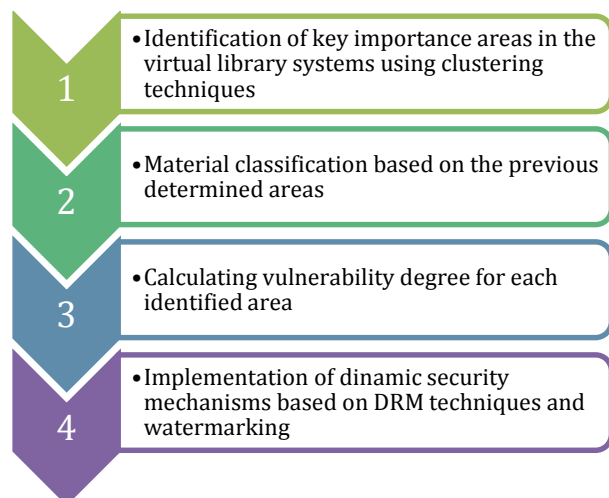


Fig. 6. Ontology implementation algorithm

Digital Rights Management access serves for digital materials protection when sharing or copying operations are executed on the content. A set of accessing rules are built in order to easily determine between the true owner and the person who's using it. DRM techniques are applied on the digital materials with the following purposes:

- protecting the digital content for unauthorized access and processing;
- safely sharing digital materials across virtual libraries.

The DRM mechanisms will action based on a

set of digital rights that will give users access upon content based on the level of vulnerability established for each domain identified with the use of ontology. The following digital rights are meant to restrict users' access:

- consulting rights – represent the right of accessing the material only for consultation purposes;
- usage rights – are the rights that allow users to include the material in their work in its original form without altering whatsoever its content compelled to cite the original source;
- processing rights – refers to the cultural dimension of the digital material that was taken from external sources and to the possibility of giving it new meaning but with the obligation of citing the original source.

6 Conclusion and Future Work

Using a wide variety of information systems for activities such as acquisition, storage, retrieval or data mining and most important sharing, a virtual library has the role to increase the academic scientific level and also to improve the quality of its provided services to the end users.

In this way a virtual library must protect its cultural heritage in its evolution through different types of systems. A cloud based library information system has other vulnerabilities than a client-server based system with regards to its content. The content is the one that must be kept safe in order that all other library services that are based upon it to work properly.

The content protection through the use of DRM techniques is not a new concern but it's definitely a challenging one. The optimization process is based upon the success of the main three directions approached: the success of the ontology implementation, the correct vulnerability classification and a good integration of the DRM controls.

Acknowledgement

"This paper is supported by the Sectorial Operational Programme Human Resources

Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/159/1.5/S/136077".

References

- [1] P. Mell and T. Grance. "The NIST definition of cloud computing." NIST, 2011
- [2] P. Pocatilu, F. Alecu and M. Vetrici, "Measuring the Efficiency of Cloud Computing for E-learning Systems," *WSEAS Transactions on Computers*, vol. 9, no. 1, pp. 42-51, 2010.
- [3] G. Garrison, S. Kim, R. Wakefield, "Success Factors for Deploying Cloud Computing," *Communications of the ACM*, vol. 55, no. 9, 2012, pp. 62-68
- [4] R. Yeluri and E. Castro-Leon, *Building the Infrastructure for Cloud Security*, Apress, 2014
- [5] S. C. Popa, M. C. Avornicului and V. P. Besfelean, "Using AMDD method for Database Design in Mobile Cloud Computing Systems," *Informatica Economica*, vol. 17 no. 1/2013, pp. 27-39
- [6] B. Williams, *The Economics of Cloud Computing*, Cisco Systems, 2012
- [7] Synergy Research Group, Microsoft Cloud Revenues Leap; Amazon is Still Way Out in Front [Online], available at: <https://www.srgresearch.com/articles/microsoft-cloud-revenues-leap-amazon-still-way-out-front>, 29 October 2014
- [8] AWS Products and Services - Global Compute, Storage, Database, Analytics, Mobile, Application, and Deployment Services, available at: http://aws.amazon.com/products/?sc_ica_campaign=ha_en_WhatIsAWS
- [9] Google Cloud Platform – Google Developers, available at: <https://developers.google.com/cloud/>
- [10] What is Microsoft Azure? – Why It's Better, available at <http://azure.microsoft.com/en-us/overview/what-is-azure/>
- [11] G. Webber-Cross, *Learning Windows Azure Mobile Services for Windows 8 and Windows Phone 8*, Packt Publishing, 2014

- [12] Enterprise Cloud Computing SaaS, PaaS, IaaS | Oracle Cloud, available at: <https://cloud.oracle.com/home>
- [13] ganeti - Cluster-based virtualization management software - Google Project Hosting [Online], available at: <https://code.google.com/p/ganeti/>
- [14] Open Source Private Cloud Software | AWS-Compatible | Eucalyptus available at: <https://www.eucalyptus.com/eucalyptus-cloud/iaas>
- [15] OpenStack Operations Guide, OpenStack Foundation, 2014
- [16] K. Jackson and C. Bunch, *OpenStack Cloud Computing Cookbook, Second Edition*, Packt Publishing, 2013
- [17] G. A. Morar, C. I. Muntean and G. C. Silaghi, "Implementing and Running a Workflow Application on Cloud Resources," *Informatica Economica*, vol. 15 no. 3/2011, pp. 15-27
- [18] K. Jeffery and B. Neidecker-Lutz (ed.), *The Future of Cloud Computing report*, [Online] available at: <http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf>, 2010
- [19] B. Wilder, *Cloud Architecture Patterns*, O'Reilly, 2012
- [20] J. van Vliet, F. Paganelli and J. Geurtsen, *Resilience and Reliability on AWS*, O'Reilly, 2012
- [21] F.G. Filip and E. Herrera-Viedma, "Big Data in the European Union," *The Bridge*, Vol. 44, No. 4, pp. 33-37, 2014, ISSN 0737-6278



Mihai DOINEA has a PhD in the field of Economic Informatics, within Academy of Economic Studies, Bucharest, Romania. His PhD thesis tackles the field of Informatics Security, with clear objectives about finding security optimization methods in distributed applications. His research is also backed up by a master diploma in Informatics Security (2008). He is an assistant professor, teaching Data Structures and Advanced Programming Languages at the Academy of Economic Studies. He published more than 30 articles in collaboration or as single author and his research interests are directed to areas such as security, distributed applications, artificial intelligence and optimization algorithms.



Paul POCATILU graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 1998. He achieved the PhD in Economics in 2003 with thesis on Software Testing Cost Assessment Models. He has published as author and co-author over 45 articles in journals and over 40 articles on national and international conferences. He is author and co-author of 10 books, (Mobile Devices Programming and Software Testing Costs are two of them). He is professor at the Department of Economic Informatics and Cybernetics within the Bucharest University of Economic Studies, Bucharest. He teaches courses, seminars and laboratories on Mobile Devices Programming, Economic Informatics, Computer Programming and Project Quality Management to graduate and postgraduate students. His current research areas are software testing, software quality, project management, and mobile application development.