Preventing Document Leakage through Active Document

Zeyad S. Aaber
School of Electronic and Computer Science
University of Southampton
Southampton, UK
zsa1g13@ecs.soton.ac.uk

Richard M. Crowder
School of Electronic and Computer Science
University of Southampton
Southampton, UK
rmc@ecs.soton.ac.uk

Nawfal F. Fadhel
School of Electronic and Computer Science
University of Southampton
Southampton, UK
nff1g08@ecs.soton.ac.uk

Gary B. Wills
School of Electronic and Computer Science
University of Southampton
Southampton, UK
gbw@ecs.soton.ac.uk

Abstract—**Electronic documents inside any enterprise environment are assets that add to the enterprise's capital in intellectual property such as design patents or customer information, securing, these assets is a priority requirement in any security system design. The security of these documents suffers when they have migrated outside the organisation security system, as there is not always a way to extend the enterprise security policy to limit/prevent access to those assets. This paper presents the challenges faced when preventing document leakage. We propose active document as a measure to control/limit access when electronic document are outside its domain.**

Keywords-Active Document; Document Leakage; Enterprise Security; Cross Domain Security.

I. INTRODUCTION

Electronic documents are an essential part of the 21th century life style. These documents contain diverse information vary from doctor appointment to bank statements. The same thing apply on cooperate businesses and governmental sector but on a massive scale. Corporations are sending document that may contain more sensitive information than a single bank account. This Information may contain Intellectual Property, financial information, pricelists, and employees or patients' private data. Losing such information has a very devastating effect on the organisation.

This paper investigates the challenges facing electronic documents leakage from private and public sector. The challenges facing electronic document are either internal or external to the domain are listed with documented legal cases. Active document structure could provide a solution to address threats facing electronic documents. This paper will start with describing the problems and the scope of the paper, then a literature review about active document in general and then specifically in a security context. We then propose using the active document concept to address these problems. Finally a discussion and future work.

II. THE PROBLEM

Enterprises and governmental bodies dedicate a considerable amount in resources to secure electronic documents inside and outside their organization structure. However, more than two thirds of the big originations in UK had a security breach in 2014 [1]. The problem is not in the lack of security precautions (security policies, security software or security mechanisms) but mostly employees' unawareness of these security precautions which contribute to 58% of the above breaches mentioned above [1]. Employee's unawareness may include losing laptop, remote access from unsafe environment, opening suspicious emails or coping data to mobile devices. So employees that have access to critical information have more devastating effect on the organisation than outsider attacker. These effects include but not limited to loss of confidential/proprietary data, reputational harm, critical system disruption and loss of current or future revenue.

Another major source of document leakage is sharing the document with third-party. "The quest for secure information sharing has been a central but elusive goal for information security for over the last three decades. The stumbling blocks are simple to understand but difficult to solve. Digital information is easy to copy and transport, and read access to any copy is as good as read access to the original" [2]. About forty percentage of the organisations did not review (or they don't know they have to) third party security procedures and about double that percentage they did not agreed on any respond plan if information security leakage occurs[3].

Document leakage is one of the continuous threads that each organization is facing no matter what size or country it belongs to. "Jonathan Pollard, who had high-level security clearance, was arrested for passing tens of thousands of pages of classified U.S. information such as satellite photographs, weapon systems data, etc., to Israel. A Libyan intelligence agent obtained the U.S. Military's officers' directory through his wife, who worked at the Department of Transportation and had access to the database of the Metropolitan Washington Council"[4]. A recent exposure of such massive governmental document is by Snowden a former employee at the National Security Agency [5];document being leaked mostly because of the employees themselves whether of their carelessness or on purpose.

III. CHALLENGES FOR DOCUMENT SECURITY

The challenges that facing document security is identified from the literature and summarised into the following points:

1. Human negligence: system users are the tools to realize the security policy for any organization. So, whatever the level of perfection the security policy reaches still there is a human negligence that causes most of document security breaches ([6]; [7]).

2. Cross-domains: different domains (military, medical and business) using different document authentication authority. So when a document travelled from one domain to another it is usually difficult to guarantee the security of the document [8]. The challenge is to produce an independent system but at the same time adaptable to accommodate diverse domain application [9].

3. Documents leakage: losing valuable information like (IP, HR information, financial and manufacturing) becomes more costly and devastating year after year ([10]; [11]).

4. Legalisation: there is deferent view for a legal value of an electronic document in the world [9]. However, they are all agreeing that the electronic document have full legal value if they can verify that the document is not Tampered with [12].

This paper proposes a method to solve the first three challenges of securing a document. Legalisation challenge is out of this paper's scope.

IV. Document Security

The basic security principles are applied which are (Confidentiality, Integrity and Availability) [13]. Document security is covered under security principles for protection of confidential information.

Confidentiality is the ability to restrict/ control access to certain information for the authorised users only. "Least Privilege" principle is used to make sure that only limited number of users whom had true need for this data is able to access it. some of the mechanisms or technologies are used to enforce information confidentiality are Access control, Authentication and encryptions [14].

Integrity is the act of assuring that the information is accurate and reliable and it is not been tempered by an authorised access or unknown entity. It mainly includes three parts; Authenticity, Accountability and Non-repudiation. Authenticity is the ability to prove that the information is not changed in an unauthorised manner. While the Non-repudiation is the ability to record every action on the information, sending and receiving, to prevent tempering and fraud but it does not guarantee the delivery of the information. Finally, the Accountability is the ability to link the user to every information action like time, access level and method used to perform the action [15].

Availability is to ensure that the information and other crucial assets are always available when needed by the user. The loss of information availability is not only when a natural disaster happens, it may because of access to information is delayed or denied due to hacker attack. This may result in business disruption and losing revenue and costumer trust. So, risks like loss of privacy, fraud, information no longer being reliable and loss of user's confidence are what security researches trying to solve [16].

V. Issues in Current Solutions

There are several solutions for securing a document inside an organisation, for instance, Document Management System DMS, Content Management System CMS and Data Loss/ Leakage Prevention DLP[6]. However, they cannot extend their protection for the document that leaves the organisation firewall or checked out (copied) by an authorised employee [6] [17].

The recent solution is the Digital Right Management DRM concept. This concept was adapted from the Entertainment domain. Basically the DRM ensures that the document is used by the authorised user even outside the organisation whilst that user is fulfilling the financial requirement for that usage. Since DRM is revenue oriented solution it sometimes reduces its control over document with less quality or appropriately out dated [2]. This discloser is not tolerated in document security case since it compromises the information confidentiality, which is one of the information security principles. Secure document sharing considers any discloser of the information as a breach and it is not accepted even with out-dated document or low resolution format.

Another issues that facing DRM as document security solution could be one or more of the following; Feasibility where some DRMs are difficult or inapplicable for the organisation IT infrastructure, Dependency some DRM depend on certain plugins and these plugins may limit the usage of the document in another domains, DRM client security is the roof of the DRM security, once the client is compromised the whole system is compromised [18].

DMS/ EDMS is limited to secure a document inside the organisation only. On the other hand it is not designed to manage cooperative work, or sharing a document and maintain it secure in multi site organisation [17].

ECM/ CMS is also limited to inside enterprise document security, once a user gains access to a document there is no way to monitor or control his/her operations on that document. Another issue is when the user keeps a local copy in his/ her machine. This copy of the document is by all mean is under the user control and there is no way to protect that document [6].

DLP is protecting the document from being sent outside the originating organisation framework by monitoring the communication channels and document archives, but it can do nothing regarding a document leaves the organisation on removable memory. The main mechanism used by DLP is to monitor the document and scanning them for a certain data. This mechanism cannot be applied on encrypted data hence there is no way to determine the actual data [6].

ERM is another way to secure a document inside the originating organisation. As ECM, a user can have an offline version of the document which the ERM can do nothing about it. Even when the document is deleted on the ERM the user still has its own version. If the document is stored in plain text then it stays as plain text, sense the ERM consider encryption as altering (tampering) for the document [6].

IRM/ ERM is the only solution has the ability to secure the document inside and outside the organisation. But this facility

comes with a price. Compatibility is one of the main issues facing IRM. Usually different venders depend on their ecosystem to provide best security exercise, so sometimes they facing issues with other vender's ecosystem (like Microsoft DRM it more powerful with MS office on Windows environment) [19]. This leads to another issue in IRM that it depends on the existing authentication mechanism, which means the IRM is secure as the authentication mechanism is. Finally having the same level of security protection to all the document is not always good strategy, doing things right is not doing the right thing [20].

VI. Active Document in the Literature

The Active Document as a concept was first introduced in 1994 [21]. The concept states that the document could be active when it has set of features more than its basic logical structure, and the "document manipulation system" in other words the presentation software has the ability to read that extra features by using some mechanisms. The active document concept emphasised on cooperative editing and authoring document such as user interface and on document indexing.

An inspiring research done In 1999 at Xerox Palo Alto Centre USA and the main focus is document management systems and document- centred collaboration [22]. Their approach based on separating the coordination information stored in the document from the actual document data that presented by the associated program. As a result the document carries its sematic within, which can be read by a middle ware they designed for that purpose. This middle ware will read this semantic information and convert them to actions on the fly (there is no need to open the file).

In 2000 Xerox introduce their prototype project "Placeless Document System" to explore the new features they propose [23]. At that point they use the phrase "Active Properties" to represent the sematic information stored inside the document by their middle ware. These active properties extend the uniform document properties and metadata so that they represent not only structure but behaviour as well. On the same time a framework for processing active documents in business domain was introduced [24]. Active document in this terminology is the document that contains both business rules and data. The framework focused on combining business rules and data for a web form at the client side and then validates the business rules before triggering any event on the database server side. However, their proposed framework works on web forms with Database management system back end in particular oracle. And they did not mention any things regarding any security aspect that may face their framework. The words "active document" used again but this time to capture user interaction with Web 2.0 applications. This time it used to add some semantic to the information available in the web. [25]

On 2012 the most recent active document concept is reintroduced in security context [26]. A new Enterprise Digital Right Management architecture was proposed to secure file being shared through the cloud among various parties (main organisation and subcontractors or outsourcing entity). The Data encapsulation is used to store the active properties which are basically security related information (access control, audit, and metadata) and "Security Kernel", which is fundamentally a piece of code [26]. The security kernel is responsible for executing security procedures. The user needs to have a "licence" file and a "Trusted Viewer" in order to read the document, the licence file explicitly describe his/ her right to access the document. While the trusted viewer could be either a lightweight viewer embedded inside the document itself or heavyweight trusted viewer using Application-programming Interface (API) to import the actual data from the document. Another way to view the document content is to export the data to eXtensible Mark-up Langue (XML) to be displayed in any regular XML viewer or any conventional viewer.

At either way of viewing the document the security kernel will accept or reject the user access, collecting and attaching metadata of the action and finally calculation new data as the action going along. After that the document stores the new data and waits to connect to a server, specialised for synchronisation propose, to synchronise the new data [26].

VII. Active Document Leakage Prevention Method

The proposed method is to make the electronic document take a part in decisions regards the basic operation performed on it weather it is inside or outside the enterprise network; in theory the electronic document should maintain its integrity by preventing content modification. The system structure is illustrated in Figure 1

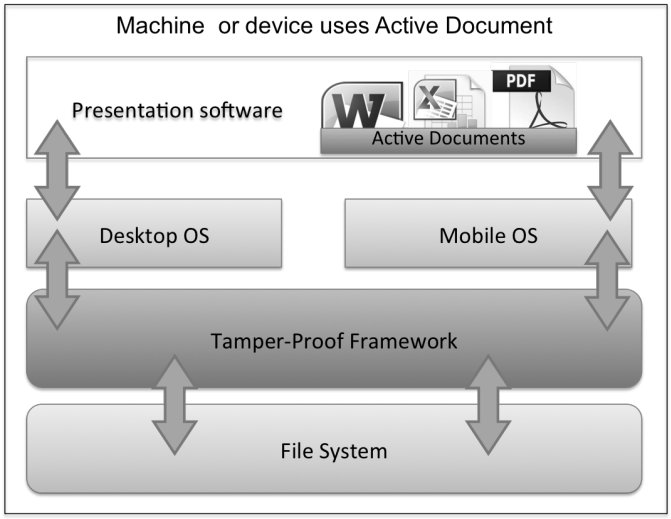

Figure 1 Active Document Leakage Prevention System

The system structure is underpinned by two concepts, the active document concept and the DRM concept as in Figure 2.

A. The active document contains the active properties, which enable the document to perform security processes.

B. The Digital/Information right Management concept provides continues security for the document inside its wrapper.

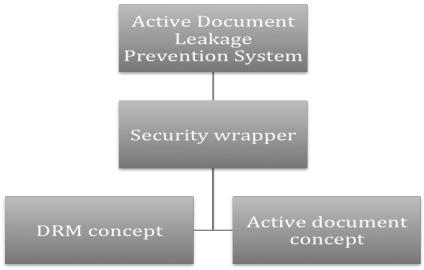

Figure 2 Tamper-proof Framework concepts

Security Wrapper is a software designed to perform a specific security tasks along with viewing a document [6].The wrapper usually used to replace the actual document presentation software in order to control the operations performed on the document (copy, past, cut, delete and print). Moreover, it used to provide secure channel to authenticate the users and enforce the organisation security policy even outside the organization firewall.

The system aims to convers any machine or device in which it installed whether it inside or outside the organisation as a wrapper for the document. This will provide two main benefits; first is that the file will be presented (viewed and edited) using the legacy presentation software (normal office suit and any PDF reader). Second, this will facilitate the system to identify each machine or device and user inside the organisation besides the privilege of hat user or machine. The framework aims to integrate seamlessly with the existing presentation software, which the user is familiar with their interface, to monitor the operation to be performed by the used on that document. By doing that the framework will enforce the security policy of the originating organisation wherever the file is resides in. Figure 3 gives an over view of the Tamper-proof framework components.

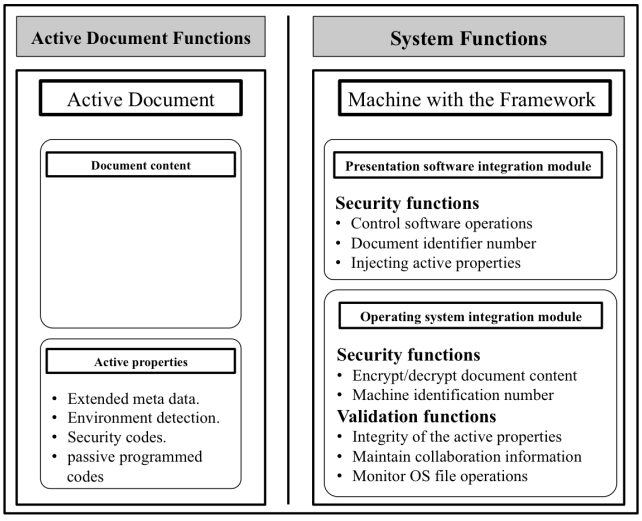

Figure 3 Tamper-proof Framework components

A. *Active Document Functions*

They focus on the following aspects; maintain extended metadata, security code for document and content verification, and finally piece of passive programming codes. The active properties are used by the system functions to enforce the originating organization security policy. Security codes are used as keys for security purposes. Passive programming codes used to complete the system functions procedures. The programming details and methods of using these keys and the passive programming codes are beyond the scope of this paper.

B. *System Functions*

These functions are the ones that responsible to provide validation and security/ control measures. Validation functions are mainly to ensure the document integrity and forcing the originating organisation security policy. The integrity here means that content of the file and the active properties of the file are not tampered with. While the security policy enforcement ensures what is the level of collaboration is allowed and who do what and when. Finally it ensures to monitor the file system activity for copying and moving content to other files.

Security functions represented by producing document identification number, machine identification number, injecting and modifying active properties inside the active document, perform encryption and decryption operation on document content and finally control the legacy presentation software operations. These functions performed by both presentation integration module and operating system integration

The modules that responsible for these functions are presentation software integration module and the operating system integration module as shown in Figure 3.

The method uses the ordinary channels to transfer the active document between domains. The ordinary channels may be emails, shared folder, or removable storage. Figure 4 Shows the overview of how the method works.

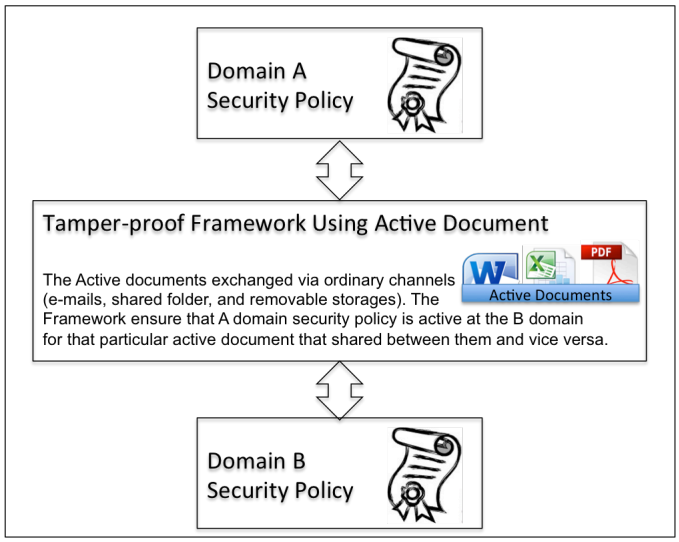

Figure 4 An overview of how the method Works

VIII. ACTIVE DOCUMENT STRUCTURE

The active document structure used in this method is adapted from the exciting work of the placeless document system [27].

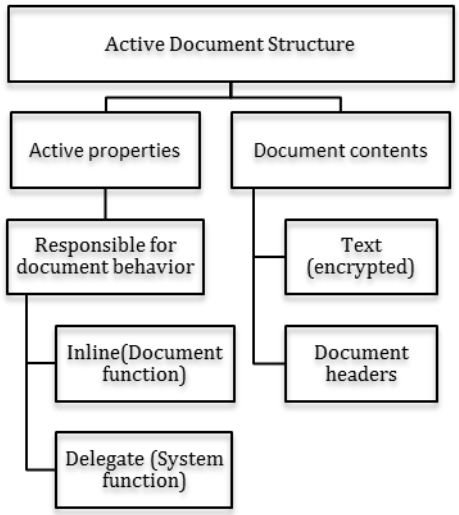

Figure 5 Active document structure

The Active document contains two main parts Active properties and Document contents as shown in Figure 5. The active properties are categorised into inline and delegates active properties. Inline active properties are the standard form of the active properties, which positioned into a document execution path to change the document behaviour via controlling document operations. These operations include but not limited to deleting content, editing content, reading content or a user who authorised to perform each operation. Each inline active properties are connected to one or more of these operations. For example an inline active property is used to monitor document activity, so the it will be positioned before reading content and writing content operation to record where, when and whom access that document content.

The delegate will be part of the System functions which shown in Figure 3. This means that system functions will provide a middle layer to all legacy presentation software on that operating system to display the document content. The active properties will make the active document user-centric instead of document centric. The system functions will read the active properties and then ensure the user can see what only permitted by these active properties. This mean the same document will be viewed differently on other user machine or on the same machine but different user is logged in.

The framework will use Java programming to ensure cross platform compatibility and all the possible domains.

IX. Conclusions and Future Work

Document leakage is one of the issues that required continuous effort to be eliminated. High amount of resources is spent in order to achieve that yet the issue is not closed. Human negligence and security policy incompatibility issues are the major challenges faces document protection. Active document concept provides the method to overcome these challenges. The active document curries its security requirement within wherever it resides whether it inside or outside its originating organisation.

Using active document as security methods to protect the document from being leaked is the new approach in this paper. The literature identifies the use of active document since 2000 till 2012 as document management and workflow facilitator. In 2012 this term was first used in security context. In this paper the active document is used as a method to solve document leakage problem not only leakage form the originating organisation but from other third parties as well so it will solve the cross domain security issue. Another feature in this method there is no need to have special kind of presentation software, instead the ordinary (legacy) presentation software will be used.

Future work will includes building full operational system after collecting more feedback on the idea and determine what is the proper metric should be used to measure the performance. After the expert review is conducted and analysed, the next step is to build a model as proof of concept. Then at that stage we can design a metrics to which the performance of the method could be measured.

Acknowledgment

The authors acknowledge the postgraduate sponsorship to Zeyad Sabah Aaber by the Higher Committee for Education Development HCED – Prime Minister Office - Iraq

References

[1] B. Innovation, “INFORMATION SECURITY BREACHES SURVEY 2014: technical report,” 2014.

[2] R. Sandhu, K. Ranganathan, and X. Zhang, “Secure information sharing enabled by trusted computing and PEI models,” *ACM Symp. Inf.*, 2006.

[3] U. S. State and C. Survey, “US cybercrime : Rising Key findings from the 2014 US State of Cybercrime Survey,” 2014.

[4] J. Park and S. Ho, “Composite role-based monitoring (CRBM) for countering insider threats,” *Intell. Secur. Informatics*, pp. 201–213, 2004.

[5] W. Macaskill and G. Dance, “NSA files decoded: Edward Snowden’s surveillance revelations explained | World news | theguardian.com,” *The Guardian*, 2013.

[6] R. Smallwood, *Safeguarding Critical E-Documents: Implementing a Program for Securing Confidential Information Assets*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2012, p. 263.

[7] I. You, L. Catuogno, A. Castiglione, and G. Cattaneo, “On asynchronous enforcement of security policies in ‘Nomadic’ storage facilities,” in *2013 IEEE International Symposium on Industrial Electronics*, 2013, pp. 1–6.

[8] M. Alawneh and I. M. Abbadi, “Preventing information leakage between collaborating organisations,” in *Proceedings of the 10th international conference on Electronic commerce - ICEC ’08*, 2008, p. 1.

[9] A. Schmidt and Z. Loebl, “Legal security for transformations of signed documents: Fundamental concepts,” *Public Key Infrastruct.*, pp. 255–270, 2005.

[10] J. ho Eom and N. uk Kim, “An Architecture of Document Control System for Blocking Information Leakage in Military Information System,” *Int. J. ...*, vol. 6, no. 2, pp. 109–115, 2012.

[11] M. A. Riley and A. Vance, “China Corporate Espionage Boom Knocks Wind Out of U.S. Companies,” *Bloomberg L.P*, 2012. [Online]. Available: http://www.bloomberg.com/news/2012-03-15/china-corporate-espionage-boom-knocks-wind-out-of-u-s-companies.html. [Accessed: 19-Jun-2014].

[12] T. Triebsees and U. M. Borghoff, “A Theory for Model-Based Transformation Applied to Computer-Supported Preservation in Digital Archives,” in *14th Annual IEEE International Conference and*

Workshops on the Engineering of Computer-Based Systems (ECBS'07), 2007, pp. 359–370.

[13] R. Mendell, *Document Security: Protecting Physical and Electronic Content*. Springfield, Illinois 62704: CHARLES C THOMAS • PUBLISHER, LTD., 2007.

[14] M. Whitman and H. Mattord, *Principles of Information Security, 4th Edition*, 4th ed. Boston, MA, USA: Course Technology, Cengage Learning, 2011, p. 656.

[15] C. Easttom, *Computer Security Fundamentals*, Second. Indianapolis, Indiana: Pearson, 2012.

[16] M. Stamp, *Information security: principles and practice*, vol. 11, no. 97. Hoboken, New Jersey: JohnWiley & Sons, Inc, 2011.

[17] R. H. Sprague, "Electronic Document Management: Challenges and Opportunities for Information Systems Managers," *MIS Q.*, vol. 19, no. 1, p. 29, Mar. 1995.

[18] S. Brook, T. Chiueh, and M. T. Road, "Display-Only File Server : A Solution against Information Theft Due to Insider Attack," pp. 31–39, 2004.

[19] EMC Corporation, "The Challenges of Deploying Information Rights Management Across the Enterprise," Bedford, MA, 2008.

[20] Manasdeep, "Information rights management implementation and challenges," Mumbai, 2012.

[21] V. Quint and I. Vatton, "Making structured documents active," *Electron. Publ.*, vol. 7, no. November 1993, pp. 55–74, 1994.

[22] A. LaMarca, W. Edwards, and P. Dourish, "Taking the work out of workflow: mechanisms for document-centered collaboration," *ECSCW'99*, no. September, pp. 12–16, 1999.

[23] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton, "Extending document management systems with user-specific active properties," *ACM Trans. Inf. Syst.*, vol. 18, no. 2, pp. 140–170, Apr. 2000.

[24] C. Nam, J. Lim, and I. Kang, "Declarative development of web applications with active documents," in *Proceedings. The 8th Russian-Korean International Symposium on Science and Technology, 2004. KORUS 2004.*, 2004, vol. 1, pp. 68–72.

[25] S. Abiteboul, P. Bourhis, and B. Marinoiu, "Satisfiability and relevance for queries over active documents," in *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '09*, 2009, p. 87.

[26] M. Munier, V. Lalanne, and M. Ricarde, "Self-Protecting Documents for Cloud Storage Security," in *2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*, 2012, pp. 1231–1238.

[27] P. Dourish, W. K. Edwards, J. Howell, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. Terry, and J. Thornton, "A programming model for active documents," in *Proceedings of the 13th annual ACM symposium on User interface software and technology - UIST '00*, 2000, vol. 2, pp. 41–50.