Towards Secure Content Based Dissemination of XML Documents

Mohammad Ashiqur Rahaman*, Henrik Plate†, Yves Roudier* and Andreas Schaad‡
* *EURECOM, 2229 route des Crêtes, 06904 Sophia Antipolis, France, {mohammad.rahaman,yves.roudier}@eurecom.fr*
† *SAP Research, 805 Avenue du Dr Maurice Donat, BP1216-06254, Mougins, France, henrik.plate@sap.com*
‡*SAP Research, Vincenz-Priessnitz-Str. 1, 76131, Karlsruhe, Germany, andreas.schaad@sap.com*

***Abstract*—Collaborating on complex XML data structures is a non-trivial task in domains such as the public sector, healthcare or engineering. Specifically, providing scalable XML content dissemination services in a selective and secure fashion is a challenging task. This paper describes a publish/subscribe middleware infrastructure to achieve a content-based dissemination of XML documents. Our approach relies on the dissemination of XML documents based on their semantics, as described by concepts that form an interoperable description of documents. This infrastructure leverages our earlier scheme [1] for protecting the integrity and confidentiality of XML content during dissemination.**

***Keywords*-XML; Ontology; Dissemination;**

I. Introduction

The advent of cross-organizational communication based on XML processing standards such as XML schema, XSL, SOAP, WSDL, or BPEL increases the number of business-related XML document exchanges through the internet. As can be verified for enterprise applications such as enterprise resource planning (ERP) or supply chain management (SCM), these documents may have a considerable size, complex structure, and rich semantics. We term such documents as 'Enterprise XML'. Today's cross-organizational communication mostly relies on a client-server interaction model which is not tailored to dissemination requirements in large organizations, in particular when the data structure may reveal internal processes. This paper introduces a publish/subscribe interaction model suitable for such business cases. Many organizations develop proprietary XML schemas to support specific application requirements like for instance, a particular data model, industrial process, or organizational structure. Such schemas may contain business critical information that needs to be protected. Very often enterprise XML is routed by untrusted intermediaries and through insecure communication channels which also asks for content confidentiality and integrity. Even though certain standards for publish/subscribe interaction exist (e.g. WS-Notification [2]), they fall short of addressing scalability together with secure content dissemination concerns.

Regarding actual service interfaces, communication parties need to agree on a certain data model (schema). Such models may evolve over time yet existing data exchanges with peers should however be maintained. We claim that, although data models may evolve or differ from one organization to another, the underlying semantics (represented in the XML documents by XML fragments) constitute a more stable and interoperable interface between organizations. Semantic web languages like RDF [3] and OWL [4] make it possible to share an ontology describing a conceptual data model, independently from XML data structures yet can be mapped to instances of XML schemas. To address security requirements, authorization policies on the semantic level, i.e. ontology, should be supported. Such a secure exchange of documents can be achieved through the separate encryption of each document node with a secret that is computed in distributed fashion by the publishers and subscribers. In this approach, an authorization on a concept triggers a secret key computation resulting into granting authorizations to multiple XML documents or portions thereof.

Previous research effort [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] targets some of the abovementioned issues, namely confidentiality and integrity of documents in a client-server environment. However, government or industry use cases that include multiple anonymous information providers and consumers require a different dissemination approach. We propose a content based dissemination of enterprise XML using a publish/subscribe infrastructure where document producers publish documents and subscribers consume those independently of each other.

II. Data Protection

Our proposal distinguishes three actors: (a) document producers who publish encrypted and encoded XML document portions that represent ontology concepts, (b) document subscribers are the end users who receive these XML document portions and (c) disseminators that form a distributed dissemination network and manage subscriptions, enforce authorization policies on behalf of the publishers, and realize the actual content delivery to subscribers. It makes it possible to deploy a publish/subscribe infrastructure based on only partially trusted XML content disseminators. The number of publishers and subscribers as well as the number and size of the XML document portions depend on the actual use case. The following security requirements must be addressed together with scalability: (1) *Confidentiality of information:* Access to documents should be limited to authorized partners, i.e. the respective publisher and authorized subscribers. This is addressed by (a) the encryption of published XML information, supported by a distributed key management and (b) an ontology-based authorization scheme that supports

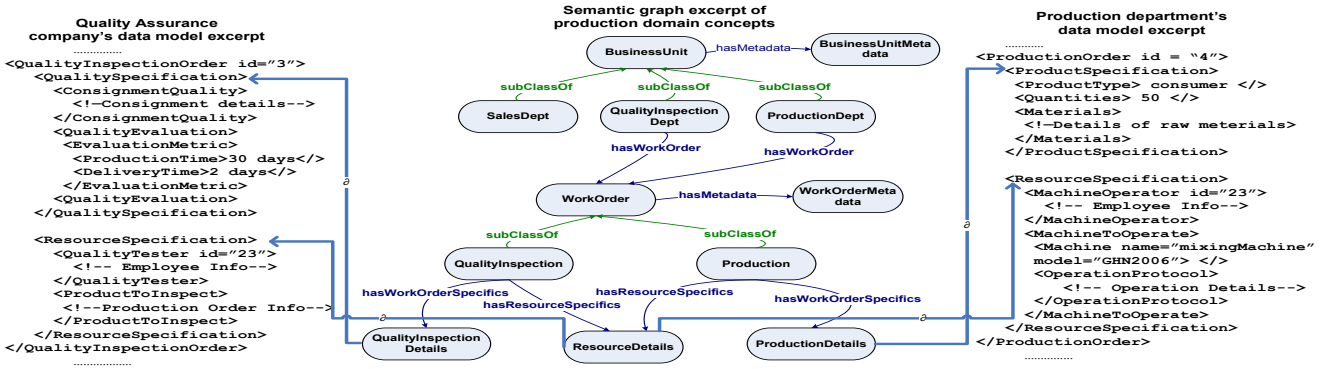

Figure 1. A semantic graph of work order document concepts in a production domain. The 'Production' and 'Quality-inspection' work order concepts are mapped to the corresponding XML data model excerpts using a mapping relation ∂.

access control and dissemination on semantic level. For (a), the EBOL-based encoding [1] of the original XML document portions ensures that the content is only readable to the authorized partners who do not need to know each other a priori. In this context, techniques developed in our prior work [16] allow a group of users to compute a group secret key independently of each other. (2) *Integrity of information:* Documents must not be altered during transit. This is achieved by an encoding method [1] that allows subscribers to verify integrity of the received documents. (3) *Confidentiality of schema information:* An information schema (e.g. XML schema) represents a valuable asset in itself (e.g. information about organizational structures or business processes can be derived from the schema) and as such needs to be confidential. This is fulfilled by using a shared ontology as the interface among organizations instead of concrete schemas. Such an ontology models all relevant business domain entities including their relationships. Every partner agrees to the ontology (Figure 1 sketches some concepts, e.g. Work Order, Production and Quality Inspection of a production process domain and their mapping to XML data models of production department and quality assurance company). The definition (or nomination) of such concepts is a prerequisite to any interaction.

Ontology-based Authorization. We describe an ontology-based authorization policy as a set of explicit rules as illustrated in Figure 2 which shows an example of a policy specified by two XML content publishers P_1 (i.e. Production department) and P_2 (i.e. Quality assurance company). R_1 and R_2 are inference rules which for instance, R_1 for the user with credential $Cred1$ is: if the user is allowed to access the concept $WorkOrder$ then he is also allowed to access to all the subclass concepts of $WorkOrder$. R_2 for the user with credential $Cred2$ is: the user is allowed to access the concept $QualityInspection$ if he has access to the concept $ResourceDetails$.

XML publisher	*User_credentials*	*Concept, C_i*	*Rule, R*
P_1	*Cred1*	*Workorder*	R_1
P_2	Cred2	*QualityInspection*	R_2

Figure 2. Ontology-based authorization policy.

III. Publish/Subscribe Model

Publish/subscribe XML content dissemination works as follows (Figure 3): (1) Prior to the first document publication, a publisher provides authorization policies to determine user authorizations which will be enforced by the dissemination network. These policies can be flexible and may evolve (see [1]). (2) An end user sends a subscription request with valid credentials (e.g. public key certificate) to a disseminator which in turn evaluates associated policies (provided by the publishers) and triggers the computation of a secret key for every group of subscribers to the same concept [16]. Unsubscription might be done on user request or be forced by the disseminator (e.g. if the user credentials expire or if authorization policies are changed). (3) The publisher of a given XML document encodes each XML document portion with its conceptual information [1], encrypts the nodes in a stipulated granularity with the computed secret key for the concept, and sends those to the disseminators. (4) Disseminators follow a dissemination protocol (detailed in [17]) in order to route the encoded XML document portions selectively to all authorized subscribers. The recipient verifies the received XML content by decoding the EBOL-based encoding, both semantically and structurally, in a verification phase which is detailed in our previous work [1].

A. Dissemination Topology

Disseminators are organized as a directed acyclic graph based on concept containment (e.g. subclass and denoted by $\preceq$) where document publishers form multiple starting roots in the dissemination network as illustrated in Figure 3. We define a $maximum\ conceptual\ block$ as the set of all concepts that are reachable by a succession of concept containment according to the ontology. Let D_i, D_j be two disseminators that disseminate two $maximum\ conceptual\ block$s represented by concepts C_i, C_j respectively. If $C_i \preceq C_j$ holds then D_i is an uplink disseminator of D_j and D_j is a downlink disseminator of D_i. Each disseminator (including publishers) maintains a distributed hash table (DHT) where the key fields and the values are the concepts and references (i.e. URL/IP) of the uplink and downlink disseminators respectively (Figure 3).

Let D_k be any disseminator reachable from D_i by following a dissemination path $D_i \rightarrow, ..., \rightarrow D_k$. C_i is the maximum conceptual block at D_i if and only if D_i or any

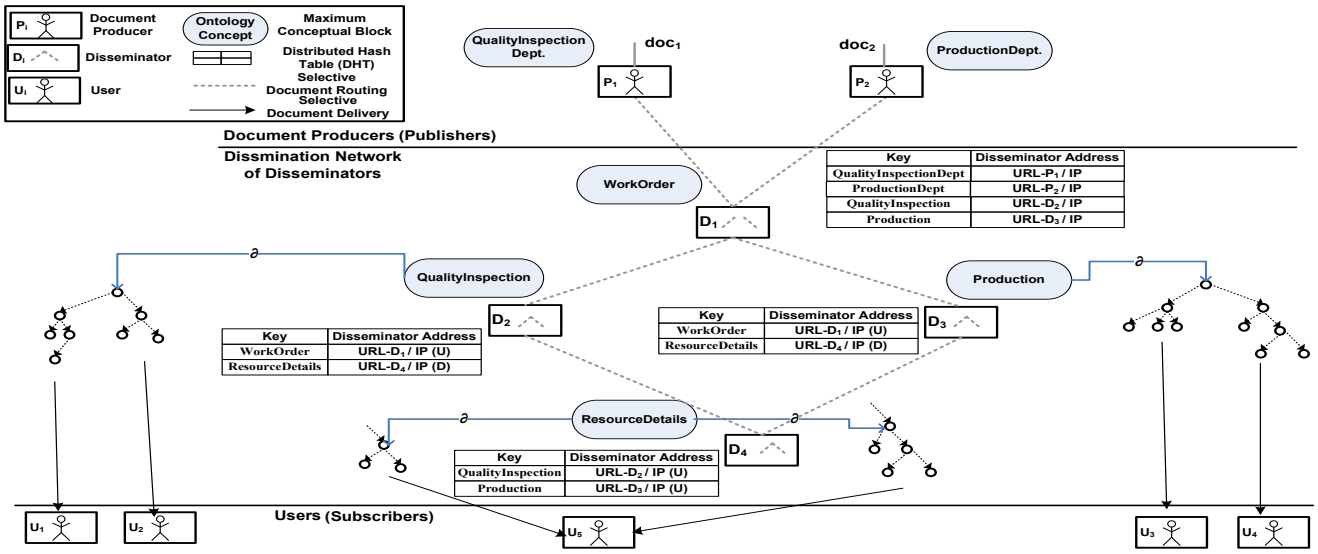

Figure 3. Publish/Subscribe infrastructure for enterprise XML content dissemination.

disseminator D_k has registered only the users who have authorizations to the concepts C_i or any of its contained concept C_j. Consequently, D_i can deliver the encoded and encrypted XML nodes to a set of subscribers such that none of them has access or has subscribed to a concept $C_m \in \mathcal{C}$, where $C_m \preceq C_i$. In effect, the disseminator D_i disseminates only the mapped XML nodes of C_i or any C_j such that $C_i \preceq C_j$. In Figure 3, the disseminator D_3 has 'Production' as the $maximum\ conceptual\ block$ for which user 3 and user 4 have collectively registered.

B. Subscription

A user u sends a subscription request (together with its credentials) for a set of concepts to a disseminator D_r. Upon receipt of that request, D_r determines the authorizations of the user u for accessing various concepts. If all authorized concepts of u (denoted by $auth\_list(u)$) are contained in the list of served concepts of the disseminator D_r, (denoted by $served\_list(D_r)$) then D_r registers the user u successfully as an authorized user and the protocol ends. Otherwise, uplink and downlink disseminators are solicited in turn. A disseminator determines a group of authorized subscribers for the same concept and sends their credentials to the publishers following the dissemination path. For each such group of subscribers, publishers trigger the independent secret key computation by sending the necessary cryptographic elements ([16]).

C. Publishing

For a new instance of a document, a publisher maps the ontology concepts to document portions and encodes and encrypts the document portions depending on their mapping and finally sends those encrypted portions to its relevant downlink disseminators.

Routing to disseminators: In addition to the DHTs and served encoded and encrypted content, each disseminator D_r also maintains the list of concepts C_k requested by other disseminators. D_r performs the following steps for each such request: (1) Determine requested concepts: find all $C_k \in served\_list(D_r)$. (2) Determine XML nodes: match concepts of step one with encrypted and encoded XML nodes. (3) Forward the encrypted and encoded XML nodes of step two to the requesting disseminator.

Delivery to users: For each subscribed user, u, the disseminator D_r performs the following steps: (1) Separate allowed concepts: find all $C_i \in auth\_list(u)$ such that $C_i \in served\_list(D_r)$. (2) Determine allowed nodes: match concepts of $auth\_list(u)$ with stored encrypted and encoded content. (3) Extract associated encrypted and encoded XML nodes. (4) Finally, send the encrypted and encoded XML nodes extracted in step three to the user u.

D. Unsubscription

We rely on a distributed computation of a group secret key by the subscribers of the same concept. The computation is performed at the subscription phase to protect the confidentiality of the XML content and of its semantics during dissemination [16]. While a new secret key should be computed by a group of subscribers in the event of a new subscriber for the same concept, the existing secret key can be used in case of a unsubscription of an existing user of a group. This is because for a successful unsubscription the relevant disseminator simply stops sending the associated XML content to that user. To unsubscribe a user u for a concept C_i, the disseminator D_r performs the following steps: (1) D_r determines the authorized XML content based on the authorizations of u for the concept C_i. (2) D_r sends a response back to u stating that unsubscription is successful and stops sending encoded and encrypted XML content to u. (3) D_r checks whether any other authorized user is subscribed for the concept C_i. If no user is subscribed, D_r forwards the unsubscription request to its uplink disseminators D_i. (4) Upon receipt of an unsubscription request for the concept C_i from its downlink disseminator, D_i sends D_r an acknowledgement and stops routing encrypted and

encoded XML content associated with C_i to D_r. D_i further checks as of step (3).

IV. Related Work

The recent years work [5, 6, 7, 8, 12, 13, 14] addresses access control issues focusing on XML structure. Their basic model is a request response paradigm in a client server architecture. Instead, this paper proposes a publish/subscribe model for semantic based dissemination.

The work of [10, 11] focuses on dissemination of XML data exploiting their hierarchical structural properties based on encrypted post order numbers. However, our proposed approach is fundamentally different as policy specification is on domain concepts and selective dissemination is performed based on the semantics captured in the concepts. The routing model of [11] is based on multi-casting of document portions from an intermediate router to the subscribers. Essentially the router may send the same document portion (i.e. subtree) multiple times to the subscriber as opposed to our approach where disseminators forbid this (see [17]).

Unlike our prior work [16] where an interested user needs to send request whenever it needs access, in this paper, we describe a publish/subscribe approach where a user can subscribe for a concept for only once and the dissemination network routes the mapped XML content to the user whenever there is a new document published.

The authors in [9, 15] propose an ontology based access control for XML documents having variant schemas and semantically related documents respectively. However, none of them considers issues related to dissemination of semantically related documents and their integrity and confidentiality.

V. Conclusion and Future Work

This paper introduces a publish/subscribe model for scalable and selective dissemination of semantically equivalent XML content to the authorized subscribers. While this model relies on a set of partially trusted disseminators to enforce access control, the confidentiality and integrity of the disseminated content is enforced using secret keys computed for concepts in a distributed fashion and special encoding methods respectively. In particular, this model ensures: (I) **Selective and scalable dissemination:** The introduction of a disseminator infrastructure and the use of ontology based policies to achieve content based dissemination ensures scalability. A publisher does not need to know a user to deliver only the nodes associated to specific concepts. At the same time, data protection mechanisms add security to this process without relying solely on that infrastructure. The topology formed by the dissemination scheme relying on the maximum conceptual blocks and DHTs is acyclic. This ensures the number of hops required for dissemination of a concept to be finite, in particular proportional to the number of successive concept containments. Being potentially at the node level, the concept authorization can be quite fine grained. Such a topology certainly facilitates efficient network usage and speedy XML dissemination compared to star, broadcast, and point to point topologies, even though communications will not be as efficient as for insecure group communication schemes. (II) **Policy evolution:** The ontology based authorization ensures the definition of flexible authorizations. Publishers may revise their existing policy, for instance the quality assurance company (i.e. P_2 of Figure 2) could add new rules: (1) dissemination of `<ResourceSpecification>` is allowed to users with $Cred2$ if `<ProductSpecification>` of production department has been disseminated already (i.e. temporal). (2) only one product, described by `<ProductSpecification>` can be tested by them (e.g. separation of duty). As access control enforcers, disseminators need an automated system for policy evaluation which we developed in [1].

We are currently implementing the dissemination method described above. We are also investigating how to extend semantics based selective document dissemination in a workflow context where receiving a document may trigger further processing of tasks and additional documents.

References

[1] M. A. Rahaman, Y. Roudier, P. Miseldine, and A. Schaad, "Ontology-based Secure XML Content Distribution," in *IFIP SEC 2009, 24th International Information Security Conference, May 18-20, 2009, Pafos, Cyprus*, May 2009.

[2] "Web Services Notification, http://www.oasis-open.org/committees/tchome.php?wgabbrev=wsn."

[3] "Resource Description Framework (RDF), http://www.w3.org/rdf/."

[4] "OWL Web Ontology Language Overview, http://www.w3.org/tr/owl-features/."

[5] W.-C. L. Bo Luo, Dongwon Lee and P. Liu, *A Flexible Framework for Architecting XML Access Control Enforcement Mechanisms*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2004, vol. Volume 3178/2004.

[6] Ernesto, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, "Fine Grained Access Control for Soap E-services," in *WWW '01: Proceedings of the 10th international conference on World Wide Web*. NY, USA: ACM, 2001, pp. 504–513.

[7] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati, "A Fine-grained Access Control System for XML Documents," *ACM Trans. Inf. Syst. Secur.*, vol. 5, no. 2, pp. 169–202, 2002.

[8] W. Fan, C.-Y. Chan, and M. Garofalakis, "Secure XML Querying With Security Views," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, NY, USA, pp. 587–598.

[9] A. Jain, D. Wijesekera, A. Singhal, and B. Thuraisingham, "Semantic-Aware Data Protection in Web Services, Proceedings of IEEE Workshop on Web Services Security held in Berkeley, CA, May 2006."

[10] A. Kundu and E. Bertino, "A new model for secure dissemination of xml content," *Systems, Man, and Cybernetics, Applications and Reviews, IEEE Transactions on*, vol. 38, no. 3, pp. 292–301, May 2008.

[11] A. Kundu and B. Elisa, "Secure Dissemination of XML Content Using Structure-based Routing," in *EDOC '06: Proceedings of the 10th IEEE International Enterprise Distributed Object Computing Conference*, Washington, DC, USA, 2006, pp. 153–164.

[12] G. Kuper, F. Massacci, and N. Rassadko, "Generalized XML Security Views," in *SACMAT '05: Proceedings of the tenth ACM symposium on Access control models and technologies*. New York, NY, USA: ACM Press, 2005, pp. 77–84.

[13] G. Miklau and D. Suciu, "Controlling Access to Published Data Using Cryptography," in *VLDB*, 2003, pp. 898–909.

[14] M. Murata, A. Tozawa, M. Kudo, and S. Hada, "XML Access Control Using Static Analysis," in *CCS '03: Proceedings of the 10th ACM conference on Computer and communications security*. New York, NY, USA: ACM Press, 2003, pp. 73–84.

[15] V. Parmar, H. Shi, and S.-S. Chen, "XML Access Control for Semantically Related XML Documents," *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pp. 10 pp.–.

[16] M. Rahaman, Y. Roudier, and A. Schaad, "Distributed Access Control For XML Document Centric Collaborations," in *EDOC '08: Proceedings of the 12th IEEE International Enterprise Distributed Object Computing Conference*, Sept. 2008, pp. 267–276.

[17] M. A. Rahaman, H. Plate, Y. Roudier, and A. Schaad, "Content Driven Secure and Selective XML Dissemination," Eurécom, Tech. Rep. RR-09-219, 05 2009.