Introducing Secure Provenance: Problems and Challenges

Ragib Hasan
Dept. of Computer Science
University of Illinois at
Urbana-Champaign
Urbana, IL
rhasan@cs.uiuc.edu

Radu Sion
Network Security and Applied
Cryptography Lab
StonyBrook University
Stony Brook, NY
sion@cs.sunysb.edu

Marianne Winslett
Dept. of Computer Science
University of Illinois at
Urbana-Champaign
Urbana, IL
winslett@cs.uiuc.edu

ABSTRACT

Data provenance summarizes the history of the ownership of the item, as well as the actions performed on it. While widely used in archives, art, and archaeology, provenance is also very important in forensics, scientific computing, and legal proceedings involving data. Significant research has been conducted in this area, yet the security and privacy issues of provenance have not been explored. In this position paper, we define the secure provenance problem and argue that it is of vital importance in numerous applications. We then discuss a select few of the issues related to ensuring the privacy and integrity of provenance information.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and Protection*; H.3.4 [**Information Systems**]: Information Storage and Retrieval—*Systems and Software*; D.4.2 [**Software**]: Operating Systems—*Storage Management*

General Terms

Security, Storage

Keywords

Provenance, Lineage, Secure storage

1. INTRODUCTION

The scientific, financial and art communities are a few of the domains where ownership and access history of objects is of paramount importance. Who owned and who accessed the objects, are vital elements considered in ascertaining their trust level.

In the case of archival records, provenance, or ownership history has been used for a long time, and has been called the *fundamental principle of archival* [16]. With the advent of financial computing systems, as well as of data-intensive scientific collaborations, the source of data items, and the computations performed during the incident processing workflows have gained increasing importance. *Provenance* (also known as information lineage) stores ownership and process history of data objects.

As an example, if Alice retrieves a document, processes a part thereof, and then forwards it over to Bob, who in turn forwards it to Charlie, the provenance of the document should contain the chain (Alice, Bob, Charlie), along with some information about the logic (e.g., process names) that operated on the document under the supervision of these users.

Provenance information has a wide-range of critical application areas. For example, scientific data processing needs to keep track of data ownership and processing workflow to ensure the trust assigned to the output data. In business environments, provenance of documents is even more important, e.g., for regulatory and legal reasons. A company's financial reports are required to contain provenance information on the path the data took during various stages of processing and the principals who performed various actions on it. For patents and other types of intellectual property litigations, tracking the ownership of documents is essential.

Although provenance of workflow and documents has been studied extensively in the past, very little work has been done on securing the provenance information. Yet, unless provenance information is secured and under the incidence of appropriate access control policies for confidentiality and privacy, it simply cannot be trusted. In this paper, we define the secure provenance problem, and explore the numerous related challenges and issues. The main contributions of this paper are to (i) identify the main challenges in trustworthy provenance, (ii) define a preliminary adversarial model, and (ii) analyze the potential security and privacy issues related to securing provenance information from the considered adversary.

The rest of the paper is organized as follows. Section 2 provides a discussion of existing literature on provenance. Section 3 presents the secure provenance problem, and the various terms associated with it. We discuss the various security and privacy issues related to trustworthy provenance in section 4. Finally, we present some applications of secure provenance in section 5.

2. RELATED WORK

Provenance has been studied extensively in the past, in particular in archival theory, and for the purpose of asserting authenticity of literary work, arts, manuscripts etc. Moreover, in recent years, provenance has also gained importance in digital realms and e-science. In [33], Simmhan et al. presented a taxonomy of the application of provenance in science, and have classified provenance systems into database-oriented, service-oriented, and miscellaneous categories. Several provenance management systems for scientific computing include Chimera [18] (physics and astronomy), myGrid [39] (biology), CMCS [29] (chemical sciences), and ESSW [19]

StorageSS'07, October 29, 2007, Alexandria, Virginia, USA.

(earth Sciences). Trio [38], based on the extended relational model named ULDB, is a system for providing lineage and provenance in databases.

Buneman et al. explored various aspects of provenance of electronic records in [11], and introduced the notions of *why-provenance* (the process that generated the data) and *where-provenance* (the origin of data) in [12]. Data provenance in curated databases is discussed in [9, 10].

Workflow provenance – the documentation of workflow in scientific computing – has been studied in [7]. The PASOA project focuses on recording workflow provenance in the grid environment [34]. Provenance information processing for data streams have been presented in [37]. The use of provenance in generating trust in social networks was explored in [21].

While most of the research activities on provenance has focused on collection, semantic analysis and dissemination of provenance, little has been done in the field of security. In [27], the authors emphasized the need for trust and provenance in information retrieval. Automated collection of provenance was discussed in [8] and Provenance Aware Storage System (PASS) – a practical implementation – was presented in [28]. However, the implementation was focused towards collection of information rather than security and trustworthiness of provenance. The Lineage File System [32] is a file system that automatically collects provenance information in the file system level. It allows rudimentary access control, where a user can set lineage metadata access flags, and the owner of a file can read all of its lineage information. However, this does not meet the various challenges for confidentiality, integrity and privacy of provenance information as discussed later in this paper.

Secure file systems have been developed to store files in untrusted servers [22]. For example, CFS [15], PAST, Farsite [6], OceanStore [25], SUNDR [26] are some systems that focus on maintaining file system integrity in a distributed environment. However, such systems focus on securing the data files rather than the chain of ownership, i.e. provenance information. While the goals are different, the techniques for detecting breach of integrity can be utilized to provide trustworthy provenance.

Ultimately, while significant research efforts have been focused on the collection, semantic analysis, and dissemination, very little has been done in securing provenance data, a vital step in achieving trust and ultimately usability of provenance as a concept.

3. SECURE PROVENANCE

Providing security and trustworthiness for provenance records is a big problem. To define the secure provenance problem, we start by informally defining a few basic terms.

Provenance: A provenance record P for a document[1] involves two components: the ownership entry for a document, and the log of the tasks applied on the document by authorized users.

Provenance chain: A provenance chain for a given document D is comprised of a time-ordered sequence of provenance records $P_1|P_2 \cdots |P_i| \cdots P_n$ of length $n > 0$, where two adjacent entries P_i and P_{i+1} indicate that user u_{i+1} obtained D from the user u_i.

Auditor: An auditor is a principal that is designated by the provenance management layer as an integrity verifier to (a given) provenance chain(s).

Secure provenance: The secure provenance problem can be informally defined as the tasks of providing assurances of integrity, confidentiality, and availability to the tasks and ownership provenance records of a provenance chain C_D for a given document D such that:

- Unauthorized parties do not have access to information stored in any of the provenance records P_i (confidentiality).
- Adversaries cannot forge a provenance record, i.e. modify content in the provenance record P_i or introduce new forged records P_{forged} in C_D without being detected (integrity).
- Authorized auditors can verify the integrity of the ownership sequence of C_D without knowing the individual records P_i within the chain (availability).
- User u_i is offered the mechanisms to selectively preserve the privacy of the provenance records pertaining to her own actions, e.g., making them available only to a selected subset of auditors (confidentiality). To avoid adversaries from masking illicit actions, certain designated auditors exist that can read any provenance chain.

In other words, secure trustworthy provenance mechanisms ensure that the integrity of provenance chains are tamper evident, their contents are confidential, and auditors can verify their authenticity without having to know the contents.

Let us illustrate with the following hypothetical scenario: *Alice creates the document D, and adds text into the document. Then she passes the document to Bob. Bob edits D and forwards it to Charlie. Charlie removes some text, and hands it to Mallory, a potentially malicious user.*

Under secure provenance assurances, unless authorized, Mallory should not be able to (a) read the provenance records for Alice, Bob, or Charlie, (b) add herself or another user into the provenance chain ($P_{alice}|P_{Bob}|P_{Charlie}$) without being detected, (c) forge Bob's provenance record without being detected, or (d) read the sequence of owners $Alice \rightarrow Bob \rightarrow Charlie$.

Now, suppose Audrey is an auditor. Audrey should be able to verify whether the D's provenance chain is indeed valid, and not tampered with. If Audrey is trusted by Alice, Bob, and Charlie to the extent that they are willing to reveal their processing task logs to her, then Audrey should be able to decode and read the encrypted provenance records for each of these users.

The above are just a few of the main assurances pertaining to securing provenance information. Next, we discuss the main challenges faced in solving the secure provenance problem.

4. CHALLENGES

Securing provenance chains requires handling several challenges. We first start by outlining our threat model and discussing the adversary of concern. Then we will explore the issue of defending against such an adversary (i.e., providing secure provenance assurances) from two angles. First, we look into the secure provenance problems by examining the confidentiality, integrity, privacy, and availability issues. Next, we look into the entire lifecycle of a provenance record, and analyze the various challenges in securing each of the phases.

[1] For illustrative purposes, we will use the term "document" to denote any data item (file, database tuple, network packet etc) under the incidence of provenance collection.

4.1 Threat model

We will consider here a malicious insider or outsider adversary with the goal of subverting the trustworthiness of provenance. Such an adversary might want to gain information about processes and actions performed on data. Some of the processes might be proprietary, and the provenance record for process/actions might reveal the information. Again, the ownership chain of a document might itself be sensitive. For example, if a document ownership chain shows that Bob, who is a secret service agent, had owned the document at some time, it might indicate that the document contains confidential information. The actions recorded in provenance may also be secret. For example, suppose Alice provides a document to Bob, who process the data using a proprietary algorithm and then hands the document to a customer Charlie. Here, Charlie should not be able to derive the proprietary algorithm by looking into the provenance record for Bob's actions.

Informally, some of the main goals of adversaries for secure provenance assurances include: gaining information from provenance records about (i) the actions performed or (ii) the ownership history of documents (e.g., to associate a user with a document), and, (iii) altering existing records or adding forged information to provenance chains (this might involve forging individual records, changing the sequence, or adding forged entries into the sequence).

We note that, without trusted hardware support, we cannot stop an adversary from copying data (manually, or electronically) to a new document, for the purpose of claiming to be its originator. The same situation occurs when a malicious adversary, with total control over his machine, can remove a provenance chain completely.

Thus, the main considered assurances given this deployment model lies rather heavily on preventing malicious entities from forging parts of the chains. Because, while an adversary will likely be able to remove document provenance chains in most deployments (in the absence of secure hardware) and claim to be their originator, it is often of little benefit in many scenarios where provenance is required. An analogy can be drawn from the art world: an art forger gains little from claiming a painting was drawn by him. Rather it is likely more advantageous to claim a painting to be a Picasso, and forge a provenance history.

4.2 Security components model

Confidentiality, integrity, and availability are main facets of general data security. To provide trustworthy provenance, we need to ensure that any proposed solution handles these facets gracefully.

4.2.1 *Integrity*

Given the nature of most of the provenance applications, integrity is perhaps the most important assurance required to achieve trustworthy provenance. The integrity of a provenance chain is again two-fold: (i) ensuring that individual provenance records are not forged, and (ii) the chain itself, that is the order of the owners of a record, is not modified.

Securing the integrity of individual records can be done by signatures, checksums, signed hashes etc. However, securing the chain is more difficult, as by definition provenance records cross multiple domain boundaries and can be passed through untrusted domains owners.

4.2.2 *Availability*

Finally, secure provenance should not reduce the availability of a document. For example, adversaries might try to invalidate a document by deleting its provenance chain. One straight-forward solution to this is to have provenance records stored in secure storage. Moreover, provenance collection should incur little or no storage and computation overheads.

4.2.3 *Confidentiality*

A provenance record chain may contain information that is of a confidential nature in different ways: (i) information about the tasks performed may be secret, and, (ii) the chain itself, that is the ownership history might contain sensitive information that should not be revealed to unauthorized parties.

Often, mere association with a document as an owner might reveal confidential and private information about a person. For example, if the name of a user appears at the provenance chain for a top secret document of the FBI, it might reveal that person's association with the FBI. Also, users might want to reveal their actions on a document only to highly trusted auditors. For example, if operations involving a proprietary algorithm are conducted on a document, and later the document is transferred to a different agency. The original user may want the auditor of that agency to be able to verify the chain, but not to reveal the information on actions taken on the document. In other words, we need a mechanism to allow selective, differentiated access control for different auditors.

4.3 Lifecycle model

We now approach the problem from a different angle, namely by exploring whether one could deploy information lifecycle mechanisms [31, 23] to achieve secure provenance assurances. Specifically, we explore associated problems in each step of the generation, storage, and transmission of provenance records.

4.3.1 *Provenance collection*

In a distributed environment, let us consider which entities would perform provenance collection tasks. In such a setting, one cannot assume a central trusted authority. This task is more likely distributed among the individual machines where data is handled. This is validated also in existing work, e.g., PASS [28] and Lineage File System [32].

Now, if we could trust all components of the organization (assuming all machines are operated by a single authority), then we could just assume trustworthiness of provenance recorded as annotations by insiders in an organizations. However, if some of the insiders are malicious, then we cannot assume all parties to record provenance information properly and require appropriate defenses.

Accordingly, one could argue that, provenance information can be collected by a (user-inaccessible) kernel process instead of a user process or application [28]. Indeed, some of the existing research activities use such techniques for recording provenance. However, we need to consider the possibility that the kernels of machines under control of malicious adversaries can be subverted, making them untrustworthy.

Ultimately, we argue that one of a few viable solutions is to benefit from trusted hardware [1, 2, 3, 4, 5] in each machine as a root of a chain of trust that ultimately ensures the trustworthiness of provenance recorded by the kernel.

4.3.2 *Storage media*

Next, let us consider the issue of storage for provenance records. In existing research a separate database [17, 32] is deployed for this purpose. However, this immediately introduces security and consistency problems. Moreover, if records are stored in the storage along with the actual documents, then we need to ensure the trustworthiness of the storage system.

4.3.3 Verification

Integrity provenance chains can be verified by auditors. However, a subset of the auditors can themselves be malicious, so we need a way to prevent them selectively. Also, some information can be more sensitive than others, and hence it should be possible to selectively say which auditors can have what level of access to a data item.

4.3.4 Migration

Yet another challenge is the handling of document migration across organizational boundaries. When a document is transferred from one organization to another, verifying signatures and obtaining associated verification secret keys, becomes a problem in the absence of a global public key infrastructure. Moreover, as part of organizational changes, document owners might leave organizations introducing new challenges in the ulterior verification of associated signatures. Delegation mechanisms need to be devised.

5. APPLICATIONS

With increased importance on the trustworthiness of electronic data records, secure provenance is becoming of essential significance. In this section, we discuss several possible applications of secure provenance.

5.1 Law

In legal circles, secure provenance information is paramount in determining the trustworthiness of electronic records such as e-mails, documents, reports, etc. Increasingly, electronic records are being used in legal proceedings. In 2006, the Federal Rules of Civil Procedure were amended to address the use of electronic information in litigations [14]. However, unless proper ownership history of a document is maintained via secure provenance, it cannot be treated as reliable evidence.

5.2 Scientific data

Provenance is already being used widely in scientific and grid computing areas to properly document workflows, data generation and processing. The source of data and the processing done on data are recorded for ensuring data quality. However, in the absence of any security mechanisms to protect the provenance information, of concern are malicious or faulty insider parties that could forge such records to fabricate data and workflow. Consequently, it is important to secure provenance chains for deployments in this domain.

5.3 Digital forensics

Similar to the legal domain, secure provenance is the essential bread and butter of digital forensics and post-incident investigation of intrusions [33, 20]. Exposure of private information due to storage security breaches has become very common in recent years [24]. This is one domain where *document* indeed can mean either a data file, a network packet or a database tuple. Interestingly if operating systems will be augmented with provenance mechanisms, we can expand the applicability of this to system files and installation kits. System binaries will now include provenance information and record all the changes made to them. If the integrity of such a provenance chain can be ensured, forensic investigators can follow and detect any changes to the binary that was introduced by the malicious intruder. Similarly, secured provenance information can be used as a certificate of authenticity in case of software distribution and updates.

5.4 Regulatory compliance

Many compliance regulations require proper documentation / audit logs for electronic records. For example, the Health Care Portability and Accountability Act (HIPAA) mandates proper logging of access and change histories for medical records [13]. A secure provenance chain is applicable here to record the modification and access history of medical records. Similarly, regulations in the financial sector, such as Sarbanes-Oxley [36], Gramm-Leach-Bliley Act [30], Securities and Exchange Commission regulation 17-a [35], etc. all require proper documentation and audit trails for financial records.

5.5 Authorship

Properly maintained provenance records can be used to resolve disputes regarding authorship and research [20]. For example, in many cases involving patents, the involved parties have to produce evidence of prior work, as well as proof of the research chronology. A secure provenance chain is ideally suited for this.

6. CONCLUSION

In this paper, we introduced and discussed the secure provenance problem, and discussed main associated challenges in achieving such desiderata. With increasing importance on trustworthy electronic record-keeping, the need for assurances of security of provenance information is more vital than ever. Ultimately we hope this paper to outline this new important vector of research and its associated challenges for researchers and practitioners in the digital community.

Acknowledgements

This research was supported by NSF Award number IIS-0331707, 0331690, CNS-0325951, and CNS-0524695. Radu Sion was supported in part through the NSF through CyberTrust Award CNS-0627554.

7. REFERENCES

[1] IBM 4758 PCI Cryptographic Coprocessor. Online at `http://www-03.ibm.com/security/cryptocards/pcicc/overview.shtml`, 2006.

[2] IBM 4764 PCI-X Cryptographic Coprocessor. Online at `http://www-03.ibm.com/security/cryptocards/pcixcc/overview.shtml`, 2007.

[3] IBM Cryptographic Hardware. Online at `http://www-03.ibm.com/security/products/`, 2007.

[4] Trusted Computing Group. Online at `https://www.trustedcomputinggroup.org/`, 2007.

[5] Trusted Platform Module (TPM) Specifications. Online at `https://www.trustedcomputinggroup.org/specs/TPM`, 2007.

[6] A. Adya, W. J. Bolosky, M. Castro, G. Cermak, R. Chaiken, J. R. Douceur, J. Howell, J. R. Lorch, M. Theimer, and R. P. Wattenhofer. Farsite: federated, available, and reliable storage for an incompletely trusted environment. *SIGOPS Oper. Syst. Rev.*, 36(SI):1–14, 2002.

[7] R. S. Barga and L. A. Digiampietri. Automatic generation of workflow provenance. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 1–9, 2006.

[8] U. Braun, S. L. Garfinkel, D. A. Holland, K.-K. Muniswamy-Reddy, and M. I. Seltzer. Issues in automatic provenance collection. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 171–183, 2006.

[9] P. Buneman, A. Chapman, and J. Cheney. Provenance management in curated databases. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 539–550, New York, NY, USA, 2006. ACM Press.

[10] P. Buneman, A. Chapman, J. Cheney, and S. Vansummeren. A provenance model for manually curated data. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 162–170, 2006.

[11] P. Buneman, S. Khanna, and W. C. Tan. Data provenance: Some basic issues. In *FST TCS 2000: Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 87–93, London, UK, 2000. Springer-Verlag.

[12] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. *Lecture Notes in Computer Science*, 1973:316–330, 2001.

[13] Centers for Medicare & Medicaid Services. The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/, 1996.

[14] U. S. Congress. Federal rules of civil procedure. Online at http://www.law.cornell.edu/rules/frcp/, 2006.

[15] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Wide-area cooperative storage with cfs. *SIGOPS Oper. Syst. Rev.*, 35(5):202–215, 2001.

[16] B. W. Dearstyne. *The archival enterprise: Modern archival principles, practices, and management techniques*. American Library Association, 1993.

[17] E. Deelman, G. Singh, M. Atkinson, A. Chervenak, N. C. Hong, C. Kesselman, S. Patil, L. Pearlman, , and M. Su. Grid-based metadata services. In *SSDBM '04: Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04)*, page 393, Washington, DC, USA, 2004. IEEE Computer Society.

[18] I. T. Foster, J. Vockler, M. Wilde, and Y. Zhao. Chimera: A virtual data system for representing, querying, and automating data derivation. In *SSDBM '02: Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pages 37–46, Washington, DC, USA, 2002. IEEE Computer Society.

[19] J. Frew and R. Bose. Earth system science workbench: A data management infrastructure for earth science products. In *SSDBM '01: Proceedings of the Thirteenth International Conference on Scientific and Statistical Database Management*, page 180, Washington, DC, USA, 2001. IEEE Computer Society.

[20] C. Goble. Position statement: Musings on provenance, workflow workflow and (semantic web) annotations for bioinformatics. In *Workshop on Data Derivation and Provenance*, Chicago, 2002.

[21] J. Golbeck. Combining provenance with trust in social networks for semantic web content filtering. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 101–108, 2006.

[22] R. Hasan, Z. Anwar, W. Yurcik, L. Brumbaugh, and R. Campbell. A survey of peer-to-peer storage techniques for distributed file systems. In *ITCC '05: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II*, pages 205–213, Washington, DC, USA, 2005. IEEE Computer Society.

[23] R. Hasan, S. Myagmar, A. J. Lee, and W. Yurcik. Toward a threat model for storage systems. In *Proceedings of the first ACM workshop on Storage security and survivability (StorageSS)*, pages 94–102, Fairfax, VA, USA, 2005. ACM Press.

[24] R. Hasan and W. Yurcik. A statistical analysis of disclosed storage security breaches. In *Proceedings of the second ACM workshop on storage security and survivability (StorageSS)*, pages 1–8, Alexandria, Virginia, USA, 2006. ACM Press.

[25] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao. Oceanstore: an architecture for global-scale persistent storage. *SIGPLAN Not.*, 35(11):190–201, 2000.

[26] J. Li, M. Krohn, D. Mazières, and D. Shasha. Secure untrusted data repository (sundr). In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, pages 9–9, Berkeley, CA, USA, 2004. USENIX Association.

[27] C. A. Lynch. When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology*, 52(1):12–17, 2001.

[28] K.-K. Muniswamy-Reddy, D. A. Holland, U. Braun, and M. I. Seltzer. Provenance-aware storage systems. In *USENIX Annual Technical Conference, General Track*, pages 43–56, 2006.

[29] J. D. Myers, T. C. Allison, S. Bittner, B. Didier, M. Frenklach, J. William H. Green, Y.-L. Ho, J. Hewson, W. Koegler, C. Lansing, D. Leahy, M. Lee, R. McCoy, M. Minkoff, S. Nijsure, G. von Laszewski, D. Montoya, C. Pancerella, R. Pinzon, W. Pitz, L. A. Rahn, B. Ruscic, K. Schuchardt, E. Stephan, A. Wagner, T. Windus, and C. Yang. A collaborative informatics infrastructure for multi-scale science. *clade*, 00:24, 2004.

[30] National Association of Insurance Commissioners. Graham-Leach-Bliley Act, 1999. `www.naic.org/GLBA`.

[31] D. Reiner, G. Press, M. Lenaghan, D. Barta, and R. Urmston. Information lifecycle management: The emc perspective. *icde*, 00:804, 2004.

[32] C. Sar and P. Cao. Lineage file system. Online at http://crypto.stanford.edu/ cao/lineage.html, January 2005.

[33] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, September 2005.

[34] M. Szomszor and L. Moreau. Recording and reasoning over data provenance in web and grid services. In *International Conference on Ontologies, Databases and Applications of SEmantics (ODBASE)*, volume 2888 of *Lecture Notes in Computer Science*, pages 603–620, Catania, Italy, 2003.

[35] The U.S. Securities and Exchange Commission. Rule 17a-3&4, 17 CFR Part 240: Electronic Storage of Broker-Dealer Records. Online at `http://edocket.access.gpo.gov/cfr_2002/aprqtr/17cfr240.17a-4.htm`, 2003.

[36] U.S. Public Law No. 107-204, 116 Stat. 745. The Public Company Accounting Reform and Investor Protection Act, 2002.

[37] N. N. Vijayakumar and B. Plale. Towards low overhead provenance tracking in near real-time stream filtering. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pages 46–54, 2006.

[38] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR '05)*, January 2005.

[39] J. Zhao, C. A. Goble, R. Stevens, and S. Bechhofer. Semantically linking and browsing provenance logs for e-science. In *ICSNW*, pages 158–176, 2004.