How Usage Control and Provenance Tracking Get Together - A Data Protection Perspective

Christoph Bier
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation
Karlsruhe, Germany
christoph.bier@iosb.fraunhofer.de

***Abstract*—These days, sensitive and personal information is used within a wide range of applications. The exchange of this information is increasingly faster and more and more unpredictable. Hence, the person concerned cannot determine what happens with his personal data after it has been released. It is highly intransparent who is accountable for data misuse. Usage control and provenance tracking are two different approaches to tackle this problem. This work compares the two concepts from a data protection perspective. The support and fulfillment of data protection requirements are analysed. Models and architectures are investigated for commonalities. Combining the two technologies can increase flexibility and effectiveness of provenance tracking and thereby enhance information accountability in practice, if resulting linkability drawbacks are properly handled. A joint architecture is proposed to support this insight.**

I. Introduction

The increasing complexity of information systems, business relations and processes demand new and innovative approaches to privacy and data protection. While the established legal perspective focuses on organizational measures and precise regulations of technical systems, novel concepts like information accountability[1] take a broader approach [2].

The European and American perspectives on the matter are traditionally different. Privacy is the term more often used in the Anglo-Saxon community. It is defined as the "right to be left alone". As a consequence, the capability to control access to ones personal data is at the heart of privacy regulation. Notice and choice, are core legal and organizational tools to obtain the right to access somebody's personal data. How the data is handled afterwards is bound to the obligations given by the a-priori notice. Data protection, the European term, focuses more on how data is dealt with. European data protection law is based on the constitutional tradition of the member states of the European Union, especially the German, which has influenced the data protection directive 95/46/EC a lot. Personal data is viewed as a good bound to social and communication relations between people, which can not be looked at out of context. Hence, there is no right to be left alone, but a balancing between the right to information self-determination of the data subject (also called "the person concerned") and the legitimate interests of the parties who want to collect and process personal data. Therefore, this work will mostly refer to the term data protection and not to privacy in the following.

While accountability is not new as a legal concept, especially in European data protection law, technologies to enforce accountability are still missing or not properly integrated. In communication relationships, somebody is accountable for a message sent, if it is detectable that he is the source of the message, that he can not claim that somebody else has sent the message and that it is provable that the message has not been changed by somebody else. The same is true for the receiver of a message. In a data protection context, the messages exchanged are personal identifiable information. But accountability is not limited to the transfer of personal data. The processing of personal data itself and the reasoning based on personal data has also to be conducted in an accountable way. But to hold somebody accountable is always a matter of transparency. If violations of privacy policies, codes of conduct or even laws are not visible or detectable, social and legal control can not oblige the person or organization accountable.

Usage control and provenance tracking are both technical concepts designed to improve information accountability. Usage control targets controlling the circumstances under which access to data is granted, but also enforces how data has to be dealt with after it has been accessed [3][4]. The handling of usage controlled data on third party systems is restricted by the policies shipped with the data, if the enforcement is guaranteed by a usage control infrastructure. Thereby, usage control extends the sphere of control of somebody into the (private) sphere of other people. This intrusion into other peoples sphere is one cause, why digital rights management (DRM), a close cousin of usage control, partially failed in practice. But in more controlled areas of application, usage control proves to be a useful tool. The protection of sensitive data on mobile business devices is such a use case [5].

Provenance tracking is traditionally used to prove the lineage of data in scientific computing [6][7]. Provenance provides information on the origins of data and the derivations of data from other data [8]. Thereby, it can be inferred who is accountable for the modification of data, how and where it happened and which other data influenced the process of creating new pieces of data. In data protection, provenance can be used to enable the data subject to carry out his right to information. In European data protection law, everybody has the right to know where the organization accountable got his data from, what the data was used for, where it was transfered to and how long it is stored. Only by knowing the exact data flow to and from the organization accountable, it can be assured that this information can be provided.

The following section will introduce the underlying models of usage control and provenance tracking (II.). Afterwards, data protection requirements are discussed (III., IV.) and the architectural aspects are taken into account (V., VI.). The presented work is rounded off by a short conclusion (VII.).

DOI 10.1109/SPW.2013.24

II. Underlying Models

A. Usage Control

In usage control, it has to be differentiated between the policy language and the actual system model. The two most well known usage control system models are $UCON_{ABC}$ by Park and Shandhu [9] and the unified system model by Pretschner et al. [10]. For the precise specification of policy languages it is referred to the relevant literature ([11][12][13][14]).

The UCON-core [3] consists of six components: subjects, objects, rights, obligations, conditions, and authorization rules. In this way, UCON differentiates between access control and the obligations which have to be enforced after access has been granted. Authorization rules are thereby divided into two classes, right-related and obligation-related. Obligation-related rules link obligations to the access grant. Subjects are the entities which hold rights on objects. Both, subjects and objects, are associated with attributes.

The unified system model defines one policy per data-subject combination. The policies are described as ECA-rules (event, condition, action) [10]. If a desired event is intercepted, the conditions have to be checked. If the conditions become true, the action is performed. In this way, mechanisms like executors, signallers, modifiers, inhibitors, and delayers can be described [15]. E.g., a simple signaller notifies the person concerned (action) each time a piece of sensitive data is opened (event), if the person concerned was not informed before (condition).

An information flow model, distinguishing between data and container, augments the mentioned system model with the additional feature of data centric policies [16]. In this model, the set of containers and the set of data are mapped to each other by a storage function, representing the current state of data stored in containers. It has to be highlighted that not only files and other static storage elements are container, but also processes, pipes and communication channels. Hence, also the transfer to a third party can be represented by a communication container.

B. Provenance Tracking

Provenance tracking models are mostly concerned with how and what kind of provenance information is stored. There are basically three kinds of provenance information: The information flow between containers (e.g., processes), information about the relationship between pieces of data, i.e., the mathematical function describing how output data is computed out of input data, and information about the environment in which the computing occurred (e.g., time constraints) [8].

The Open Provenance Model aims to standardize the exchange of provenance information [17]. The three basic components of this model are agents, artifacts, and processes. Because provenance is data driven and not event driven like usage control, only agents can be mapped one-to-one on usage control subjects. An artifact is the representation of a piece of data at a given moment in time. It reflects a relationship between a datum and a container. In this way, it is possible to describe chains of predecessors and successors in a data flow. As containers and data are conceptual time independent pre-defined sets with distinct elements, a predecessor-successor relationship can not exist between them. The Open Provenance Model is now evolved into a W3C recommendation [18].

TABLE I. Fullfillment (F) of, Support (S) of and Issues (I) with Data Protection Targets

Data Protection Target	Usage Control	Provenance	Combined
Confidentiality	S	(F)	S, (F)
Integrity	(F)	S, (F)	S, (F)
Availability	(S), (F)	(F)	(S), (F)
Unlinkability	S, (I)	I	S, (I)
Transparency	(S), (F)	S	S, (F)
Intervenability	S, (I)	(F)	S, F

III. Data Protection Requirements Satisfied by Usage Control and Provenance Tracking

The usefulness of usage control and provenance tracking depends on the fulfillment and support of data protection requirements. The following section describes such requirements and analyses the relationship between the requirements, usage control, and provenance tracking (see also Table I).

A. Data Protection Requirements

A wide range of fine grained requirements can be derived from data protection and privacy law, international standards and guidelines, statements of information protection officers, and publications of interest groups and researchers. The most well known guiding rules for data protection and privacy are the OECD guidelines on data protection [19]. But because they are a compromise between different privacy cultures, they lack structure and delimitation. The data protection targets developed by Pfitzmann and Rost follow a more systematic approach [20]. They combine the traditional security requirements of confidentiality, integrity, and availability with the three distinct privacy specific requirements of unlinkablity, intervenability and transparency [21]. The requirements are organized in pairs (unlinkablity/transparency, confidentialiy/availability, integrity/intervenability), reflecting the need to balance them against each other. It does not however follow that balancing privacy is a zero-sum game. The requirements should not be understood in a narrowly defined way, but as broad principles enclosing other fine grained privacy requirements. E.g., the minimization of data collection has not to be listed separately as it is a measure to enhance unlinkability and confidentiality. Confidentiality is improved by data minimization as unknown data is not accessible for anybody. Likewise, purpose binding is a measure to enhance unlinkablity. In this broad sense, accountability is part of the transparency requirement.

B. Usage Control

In general, related policies to all six data protection requirements can be defined in usage control policy languages like OSL [12]. Unlinkability can be targeted by forbidding that two pieces of data are copied to the same container (e.g., file). Data retention policies like “delete my data after 30 days” support unlinkability as well. Data retention policies imply that the amount of data stored is reduced after retention. Thereby, they minimize data storage form a long-term perspective.

Inhibition of actions (e.g., “send data is not allowed”, “Person X is not allowed to open file Y”) proliferates confidentiality. Notifying mechanisms translate transparency into policies. Intervenability (e.g., asking for consent) can also be described in policies. But notifying the person concerned or asking for consent requires additional infrastructure independent from the modification or inhibition of the actual action happened. Availability can partially be supported by inhibiting the deletion of required data.

Besides the wide range of requirements supported by usage control policies, it depends on the actual implementation and deployment of an usage control infrastructure which requirements are fulfilled or violated.

C. Provenance Tracking

The fulfillment of the data protection requirement “transparency by default” is the major goal of provenance tracking. The verification of the handling of personal data, and, hence, the continuous ensurance of transparency, has an external and an internal component. External transparency is primarily reached by providing comprehensive information serving the right to access respectively the right to information. Internal transparency is based on an internal monitoring and auditing system. For both types of transparency, provenance data is necessary. Regarding the integrity requirement, provenance data can also be used to verify the integrity of the underlying data, if the integrity of the provenance data itself is guaranteed.

It depends on the actual use case, which information about the lineage of data has to be collected. Given the right to information, only the external sources and sinks of data, and where and for which purpose data is stored, is necessary. In other cases, especially for internal monitoring, detailed process information may be required.

Provenance tracking is not suitable to support the other data protection requirements mentioned in section III-A. Nevertheless, these requirements have to be respected when deploying provenance tracking in real systems. This issue will be discussed in the following section.

IV. Issues Resulting from the Application of Usage Control and Provenance Tracking

Usage control and provenance tracking do not only solve data protection issues, but also create new ones. Especially the requirement of data minimization is violated by provenance tracking, and, partially, by usage control. Each additional information flow trace recorded by provenance tracking broadens the knowledge about the behavior of the individuals concerned with the processing of personal information. The connection of provenance traces via a data subject ID, necessary for allowing the access to provenance information per data subject (right to information), simplifies the linkability of sensitive data. Thereby, provenance is an example for a possible conflict between transparency and unlinkability. Usage control can also link different data traces if an information flow model is integrated and an identifier of the data subject is part of the policy. Linkability is not only a problem for data protection, but can also violate trade secrets [22]. Relationships with other business partners are a competitive advantage of an enterprise.

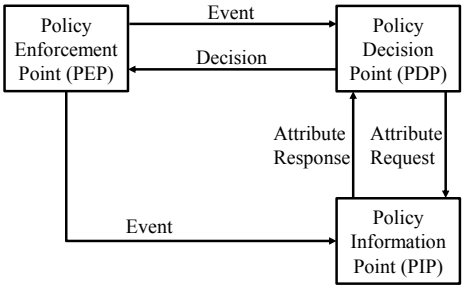

Fig. 1. Usage Control Architecture

Usage control has larger problems with intervenability. If policies can not be changed ex-post, the free will of the data subject can not be expressed. The integrity, confidentiality and availability of usage control systems and provenance data have to be addressed as well [23]. If it can not be guaranteed that a usage control system is up and running, it can be circumvented like it happened to DRM-systems many times before. If provenance information can be altered ex-post, the lineage of data and the violation of obligations can not be proven. Access rules to provenance data are different from the access to data itself [24]. A distributed usage control and provenance tracking infrastructure can not work if the peers are not available. In the worst case usage of data has to be denied by default.

Less intrusive but also less obliging are P3P [25] for notice and consent and HTTPA [26], an HTTP based protocol allowing the negotiation of obligations and the tracking of information flows in between different parties over the web [27]. P3P has failed to be accepted in wildlife since over a decade, but the more recent HTTPA introduces more flexible concepts and could, if the end user is well supported by appropriate systems, change the game.

V. Components Composing a Usage Control and Provenance Architecture

There is no one best solution to design a usage control and provenance architecture. But there are a few distinct components, each architecture for usage control and each architecture for provenance must have. After describing the essential components for usage control and provenance systems, the combination of both is motivated in the next section.

Usage control consists of three basic aspects (see Figure 1). First of all, there must be a way to detect events happening in the system usage control policies have to be enforced on. The place where an event is detected is also the best place to intercept, modify and inhibit it. Hence, the first fundamental component of usage control is the so called policy enforcement point (PEP) [11]. If an event is intercepted, it has to be decided if the event itself has an influence on the state of mechanisms reflecting a policy and if the event has to be modified based on the rules of a policy. The component responsible therefor is the policy decision point (PDP). As it makes no sense to collect information about attributes of subjects and objects as well as the global information flow state for each policy in the PDP, the third component, the policy information point (PIP) has been introduced [16]. It adds a distinct conceptual

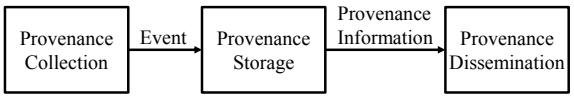

Fig. 2. Provenance Tracking Architecture

new aspect by making information flow tracking available for policy decisions. Because of that, data centric instead of container centric policies are possible.

The actual storage of policies can be part of the PDP or part of a different component like a PSP (Policy Storage Point). But from a conceptional point of view, there is no difference. Adding components for policy management and distribution justifies no significantly new conceptual perspective either. Further obligations, which can not be enforced in the same system component as the PEP is located in, may create needs for an additional Policy Execution Point (PXP) or obligation service. But as the functionality is very similar to the one of the PEP, the PXP can also be omitted in the general architecture. Hence, the three-tier-architecture of PEP, PDP and PIP represents all major ideas of usage control.

The general provenance architecture also consists of three components (see Figure 2). The first component for provenance collection (also called recording probe) is very similar to the PEP [28][29]. The main difference is that provenance collectors do not modify or inhibit the events detected. In many cases, to change the chain of cause and effect would violate the integrity of the system. Lineage has to be observer-independent.

As it is a key feature, what kind of provenance is stored and in which way the lineage of a system is modelled, this requires another component, the provenance storage component [8][30][31]. The events detected by the provenance collection component are described with the relevant attributes and forwarded to the storage component to be processed there. The provenance dissemination component refers to the ability to query and distribute provenance information. It also includes the user interface for accessing provenance.

Provenance collection, storage and dissemination can be augmented by distinct aggregators [28]. But these aggregators can either be viewed as post-processing components outside of the scope of actual provenance tracking or as a factor for moving provenance up the distribution ladder.

Implementation wise, there are three different ways to deploy a PEP or provenance collector on a system layer or application. Firstly, the component could be integrated into the application to be monitored. In this case, which is highly intrusive, the application has to be modified. Provenance architectures for scientific computing have the advantage, that most scientific computing applications already provide logging functionality which is sometimes as powerful as the collector has to be. The other two possibilities are adding an adapter for an interface of the application or writing a wrapper encapsulating the whole application. Adapters are designed from the perspective of a parent process, controlling the information flow in between these two processes. Wrappers are build around the application itself [28]. While adapters and wrappers are easier to deploy, usage control relies for obligations other than inhibiting or modifying the information flow from and to the application on monitoring the actual processing of data in an application itself. For provenance tracking, wrappers and adapters can be sufficient, if the information flow in an application is negligible and process information (e.g., how data is combined) is not necessary.

VI. Combination of Usage Control and Provenance Tracking

If combined, usage control and provenance tracking can complement each other. Regarding functionality, provenance tracking enhances the ability of usage control to notify the person concerned on what happens with his data. Tracking sensitive information independent from pre-defined policies also allows to deploy policies relating to already distributed data later on. On the other hand, provenance tracking can benefit from explicit usage control policies by differentiating between different depths of tracking and by allowing to intervene into data processing ex-post [28].

From an architectural point of view, the components partly overlap and partly complement each other (see Figure 3). While a potent PEP can also carry out provenance collection, it is not the case for the PIP and provenance storage. They are similar in functionality, but different in detail. In general, usage control needs only to determine the current state of the relationship between data and container, while for provenance also the relationship between successor, current state and predecessors of a data representation are important. In usage control, the history of a system directly influences the state of the mechanisms representing the policies.

Performance wise, fully integrating provenance storage and PIP would result in serious issues. If the precise relationship between data and container necessary for usage control is stored in a history, each system call would induce write operations on the database. Hence, provenance should rely on more abstract concepts of relationships between data and containers if detailed information is not needed. This is particularly true regarding the right to information. In this case, only present storage containers, but past transfers of data have to be known.

VII. Conclusion

Usage control and provenance tracking have more in common than it seems at first glance. The underlying models complement each other and the implementations of some components can be reused for the purpose of the respective other technology. Combining usage control and provenance could bring transparency and information accountability one step forward, but also creates some new data protection issues. Lineage data, combined with knowledge on the current information flow state, increases the threat to conclude to a persons behavior by linking personal data. But the benefits should outweigh the drawbacks, if the different aspects are wisely balanced. It is a valuable goal to develop a provenance tracking infrastructure based on usage control technologies in future work, adding boundaries to reduce the problem of linkability.

Acknowledgment

This work was funded by Fraunhofer Gesellschaft Internal Programs, Attract 692166.

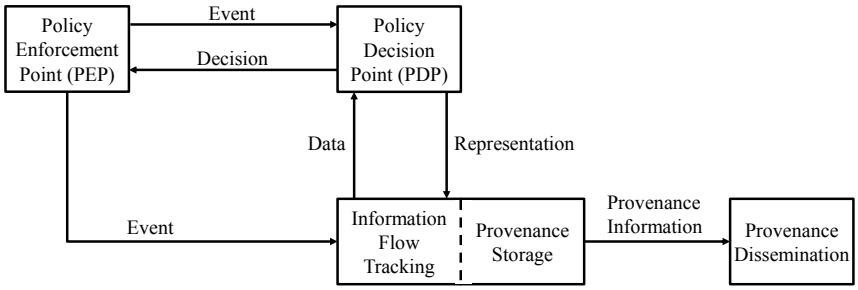

Fig. 3. Combined Usage Control and Provenance Tracking Architecture

REFERENCES

[1] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information Accountability," *Communications of the ACM*, vol. 51, no. 6, pp. 82–87, Oct. 2008.

[2] L. Kagal and H. Abelson, "Access control is an inadequate framework for privacy protection," in *Proceedings of the W3C Privacy Workshop*, no. June, 2010, pp. 1–6.

[3] J. Park and R. , "Towards usage control models: beyond traditional access control," in *Proceedings of the seventh ACM symposium on Access control models and technologies (SACMAT '02)*. Monterey, CA: ACM, 2002, pp. 57–64.

[4] A. Pretschner, M. Hilty, F. Schutz, C. Schaefer, and T. Walter, "Usage control enforcement: Present and future," *IEEE Security & Privacy*, vol. 6, no. 4, pp. 44–53, 2008.

[5] D. Feth and A. Pretschner, "Flexible data-driven security for android," in *IEEE Sixth International Conference on Software Security and Reliability (SERE)*, june 2012, pp. 41 –50.

[6] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science," *ACM Sigmod Record*, vol. 34, no. 3, pp. 31–36, 2005.

[7] J. Freire, D. Koop, E. Santos, and C. T. Silva, "Provenance for Computational Tasks: A Survey," *Computing in Science & Engineering*, vol. 10, no. 3, pp. 11–21, 2008.

[8] L. Moreau, P. Groth, S. Miles, J. Vazquez-Salceda, J. Ibbotson, S. Jiang, S. Munroe, O. Rana, A. Schreiber, V. Tan, and L. Varga, "The provenance of electronic data," *Communications of the ACM*, vol. 51, no. 4, pp. 52–58, 2008.

[9] J. Park and R. Sandhu, "The UCON ABC usage control model," *ACM Transactions on Information and System Security*, vol. 7, no. 1, pp. 128–174, Feb. 2004.

[10] A. Pretschner, M. Büchler, M. Harvan, C. Schaefer, and T. Walter, "Usage Control Enforcement with Data Flow Tracking for X11," in *Proceedings of the 5th International Workshop on Security and Trust Management (STM*, Saint Malo, 2009, pp. 124–137.

[11] B. Katt, X. Zhang, R. Breu, M. Hafner, and J.-P. Seifert, "A general obligation model and continuity: enhanced policy enforcement engine for usage control," *Proceedings of the 13th ACM symposium on Access control models and technologies (SACMAT '08)*, pp. 123–132, 2008.

[12] M. Hilty, A. Pretschner, D. Basin, C. Schaefer, and T. Walter, "A policy language for distributed usage control," in *Computer SecurityESORICS 2007*. Springer, 2007, pp. 531–546.

[13] OASIS, "eXtensible Access Control Markup Language (XACML)," 2005.

[14] R. Pucella and V. Weissman, "A formal foundation for ODRL," in *Workshop on Issues in the Theory of Security (WITS)*, Barcelona, 2004.

[15] A. Pretschner, M. Hilty, D. Basin, C. Schaefer, and T. Walter, "Mechanisms for Usage Control," in *Proceedings of the 2008 ACM symposium on Information, computer and communications security (ASIACCS '08)*. Tokyo: ACM, 2008, pp. 240–244.

[16] A. Pretschner, E. Lovat, and M. Büchler, "Representation-independent data usage control," *Proceedings of the 6th international conference, and 4th international conference on Data Privacy Management and Autonomous Spontaneus Security*, pp. 122–140, 2012.

[17] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. L. Simmhan, E. Stephan, J. V. den Bussche, and J. Van den Bussche, "The Open Provenance Model Core Specification (v1.1)," *Future Generation Computer Systems*, vol. 27, no. 6, pp. 743–756, Jun. 2010.

[18] World Wide Web Consortium, "An Overview of the PROV Family of Documents," 2012, available online at http://www.w3.org/TR/prov-overview/, visited on February 10th 2013.

[19] OECD, *Privacy Online: OECD Guidance on Policy and Practice*. OECD Publishing, 2003.

[20] M. Rost and A. Pfitzmann, "Datenschutz-Schutzziele revisited," *DuD*, vol. 33, no. 6, pp. 353–358, 2009.

[21] M. Hansen, "Top 10 mistakes in system design from a privacy perspective and privacy protection goals," in *Privacy and Identity Management for Life*, ser. IFIP Advances in Information and Communication Technology, J. Camenisch, B. Crispo, S. Fischer-Hbner, R. Leenes, and G. Russello, Eds. Springer, 2012, vol. 375, pp. 14–31.

[22] R. Herkenhöner, H. de Meer, M. Jensen, and H. C. Pöhls, "Towards Automated Processing of the Right of Access in Inter-organizational Web Service Compositions," in *Proceedings of the 6th World Congress on Services*, Jul. 2010, pp. 645–652.

[23] S. Xu, Q. Ni, E. Bertino, and R. Sandhu, "A Characterization of the problem of secure provenance management," in *IEEE International Conference on Intelligence and Security Informatics (ISI'09)*. Dallas, TX: IEEE, 2009, pp. 310–314.

[24] U. Braun, A. Shinnar, and M. Seltzer, "Securing Provenance," in *Proceedings of the 3rd conference on Hot topics in security (HOTSEC'08)*, Berkeley, CA, 2008, pp. 1–5.

[25] World Wide Web Consortium, "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification," 2002, available online at http://www.w3.org/TR/P3P/, visited on February 10th 2013.

[26] O. Seneviratne and L. Kagal, "Usage Restriction Management for Accountable Data Transfer on the Web," in *IEEE International Symposium on Policies for Distributed Systems and Networks (IEEE Policy 2011)*, 2011.

[27] O. Seneviratne, "Augmenting the web with accountability," in *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012, pp. 185–190.

[28] F. Curbera, Y. Doganata, A. Martens, N. Mukhi, and A. Slominski, "Business provenance–a technology to increase traceability of end-to-end operations," *On the Move to Meaningful Internet Systems: OTM 2008*, pp. 100–119, 2008.

[29] B. Demsky, "Garm: Cross application data provenance and policy enforcement," in *Proceedings of the 4th USENIX conference on Hot topics in security (HotSec'09)*. Montreal: USENIX Association, 2009, p. 10.

[30] K.-K. Muniswamy-Reddy, U. Braun, D. A. Holland, P. Macko, D. Maclean, D. Margo, M. Seltzer, and R. Smogor, "Layering in Provenance Systems," in *Proceedings of the 2009 USENIX Annual Technical Conference (USENIX '09)*, Berkeley, CA, 2009.

[31] M. Seltzer, D. A. Holland, U. Braun, and J. Ledlie, "Provenance-Aware Storage Systems," Harvard University, Cambridge, MA, Tech. Rep., 2006.