CERIAS Tech Report 2004-47

A ROADMAP FOR COMPREHENSIVE ONLINE PRIVACY POLICY

by Annie I Anton, Elisa Bertino, Ninghui Li, Ting Yu

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

A Roadmap For Comprehensive Online Privacy Policy Management

Annie I. Antón (*), Elisa Bertino (**), Ninghui Li (**), Ting Yu
(*) CS Department, North Carolina State University
e-mail: anton@csc.nscu.edu, yu@csc.nscu.edu
(**) CERIAS and CS Department, Purdue University
e-mail: bertino@cerias.purdue.edu, ninghui@cs.purdue.edu

Information technology advances are making Internet and Web-based system use the common choice in many application domains, ranging from business to healthcare to scientific collaboration and distance learning. However, adoption is slowed by well-founded concerns about privacy, especially given that data collected about individuals is being combined with information from other sources and analyzed by means of powerful tools (i.e., data mining tools). Effective solutions for privacy protection are of interest to industry, government and society at large, but the challenge is to satisfy the often-conflicting requirements of all these stakeholders. Enterprises need mechanisms to ensure that their systems are compliant with both the policies they articulate and law. Moreover, they need to understand how to specify, deploy, communicate and enforce privacy policies. Legislators and regulatory bodies need mechanisms to verify how privacy-related laws are actually enforced by enterprises in their software systems. Finally, end-users must be able to easily understand privacy policies [AEB04] and need effective, transparent and comprehensible online privacy-protection mechanisms.

Significant efforts in industry are seeking to better protect sensitive information online and better communicate the mechanisms used to do so in the form of privacy policies. However, existing solutions are still fragmented and far from satisfactory. For example, existing languages for specifying privacy policies lack a formal and unambiguous semantics, are limited in expressive power, and lack enforcement and auditing support [LYA03]. End-user privacy management tools are limited in capability or difficult to use. To provide effective online privacy protection, a comprehensive framework that covers the entire privacy policy life-cycle is needed. This life-cycle includes enterprise policy creation, enforcement, analysis and auditing, as well as end-user agent presentation and privacy policy processing. Trustworthy privacy protection can only be attained when broad consideration is given not only to IT solutions, but also to a wide range of perspectives from other disciplines. To this end, technical attempts to support privacy policy management must take into account the human, legal and economic perspectives that are relevant to privacy.

In this paper, we present a comprehensive architectural framework that supports the privacy policy life-cycle. We identify the relevant technological and non-technical components required to support this life-cycle, showing the relationships between these components. The framework suggests a detailed roadmap for research to be undertaken before sound privacy solutions may be realized.

Privacy Policy Technologies

To make privacy policies more readable and enforceable, two privacy policy specification languages have emerged, P3P and EPAL as we now discuss.

Platform for Privacy Preferences (P3P) Project

The W3C's Platform for Privacy Preferences (P3P) Project [P3P, Cran02, Mar02] enables websites to encode their data-collection and data-use practices in a machine-readable XML format, known as P3P policies [Mar02]. The W3C has also designed APPEL (A P3P Preference Exchange Language) [Lang02], which allows users to specify their privacy preferences. Ideally, through the use of P3P and APPEL, a user agent (a program working on the user's behalf) should be able to check a Website's privacy policy against the user's privacy preferences, and automatically determine whether the Website's data-collection and data-usage practices are acceptable to the user. P3P appears to be the most widely used (if not the only) language for encoding enterprises' privacy policies for consumption by end-users. However, P3P has several limitations and shortcomings that need to be addressed.

The P3P language does not have a clear semantics and can therefore be interpreted and presented differently by different user agents. Companies may be reluctant to provide P3P policies on their websites, because policies may be misrepresented. Quoting from CitiGroup's position paper [Sch02], "The same P3P policy could be represented to users in ways that may be counter to each other as well as to the intent of the site." "... This results in legal and media risk for companies implementing P3P that needs to be addressed and resolved if P3P is to fulfill a very important need." Furthermore, a policy specified in P3P may be internally inconsistent [LYA03].

The fundamental reason underlying the aforementioned technical difficulties is that the need for a semantics was apparently overlooked in the initial design of P3P, leaving too much freedom for user agents to misinterpret P3P policies. As discussed in [LYA03], the problem is not just about the ambiguity of vocabularies in P3P, but also about how the different components (i.e., collected data items, purposes, recipients and retentions) in a P3P statement interact. Additionally, the expressive power of P3P is limited [HJW03, Sch02, SHW02]. Many statements in a natural language privacy policy cannot be expressed in P3P, including, for example, how long data will be stored, what security mechanisms are in place to protect stored data, and what kinds of data are not collected or shared, etc.

Though Websites are starting to post their P3P policies, the majority of online privacy policies are published in natural language. Currently, only textual policies are legally binding for an enterprise. Natural-language privacy policies cover a much broader scope of an enterprise's privacy practices than P3P policies. Moreover, natural-language policies tend to be more ambiguous and incomplete [AEB04], making it difficult to maintain consistency between natural-language policies and their more formal machine-readable representations. Tools are needed for translating natural-language policies into machine readable and enforceable policies to facilitate consistency checking. Policy translation tools will enable large-scale processing of textual privacy policies and increase general understanding about the current state of privacy practices.

The P3P framework does not address enforcement or auditing. Currently, an enterprise has no way to determine whether their published privacy policy is actually enforced within their information systems; nor can it prove to other parties that adequate procedures have been followed to ensure compliance with its privacy policy. This problem is exacerbated by the fact that an enterprise shares customer data with other business partners, which may have different privacy practices [AHB04]. Even within a single organization, multiple privacy policies often exist [AEB04]. Tools are thus needed for comparing and analyzing different privacy policies, and to enforce privacy-aware information flow to thwart inappropriate information flows [AHB04].

Enterprise Privacy Policy Enforcement

Researchers at IBM are developing enterprise privacy architecture solutions [KSW02]. Karjoth et al. [AHK03, KSW02] proposed a privacy-centric access control language (E-P3P and its successor EPAL). EPAL (Enterprise Privacy Authorization Language) [AHK03] is an abstract-level access control language, with features devoted to privacy protection, e.g., data accessing purposes. We identify the following limitations of existing work.

First, the efficient and correct enforcement of policies specified in EPAL (or in a language for similar purposes) in the data storage layer has not been addressed. Policies specified at the EPAL level need to be enforced at the time data is accessed. In most cases, such data is stored in databases and is accessed frequently. Thus, if every data access had to rely on external policy evaluation, the performance would be unacceptable.

Second, the relationship between policies at the P3P level and the EPAL level has not been adequately addressed. Karjoth et al. [KSH03] proposed to generate P3P policies from EPAL policies. We disagree with this approach. Privacy policies represent long-term promises made by an enterprise to its end-users and are determined by business practice and legal concerns. On the other hand, access control policies represent internal data handling practices that may change more frequently. It is undesirable to change an enterprise's promises to customers every time an internal access control rule changes. In fact, a privacy enforcement mechanism should be able to grandfather data and associated policies (to limit scope of impact when policies change).

Third, EPAL does not address situations arising from information flows between applications under different privacy policies. The sticky policy paradigm [KSW02], which associates relevant consents with users' data so that they can be enforced during access control decisions, can help to a certain extent. However, most data exchange interfaces today do not support sticky policies; theory and tools to control information flows to other applications governed by different privacy policies are needed to ensure that the correct privacy policy is enforced.

A Comprehensive Framework for Online Privacy Protection

We now provide a general overview of the framework's key components and desirable functionalities and interactions. Figure 1 shows the architectural representation of a framework for privacy policy management.

Enterprise Side: To support the complete life-cycle of a privacy policy, the framework's enterprise side is organized according to a three-tier model.

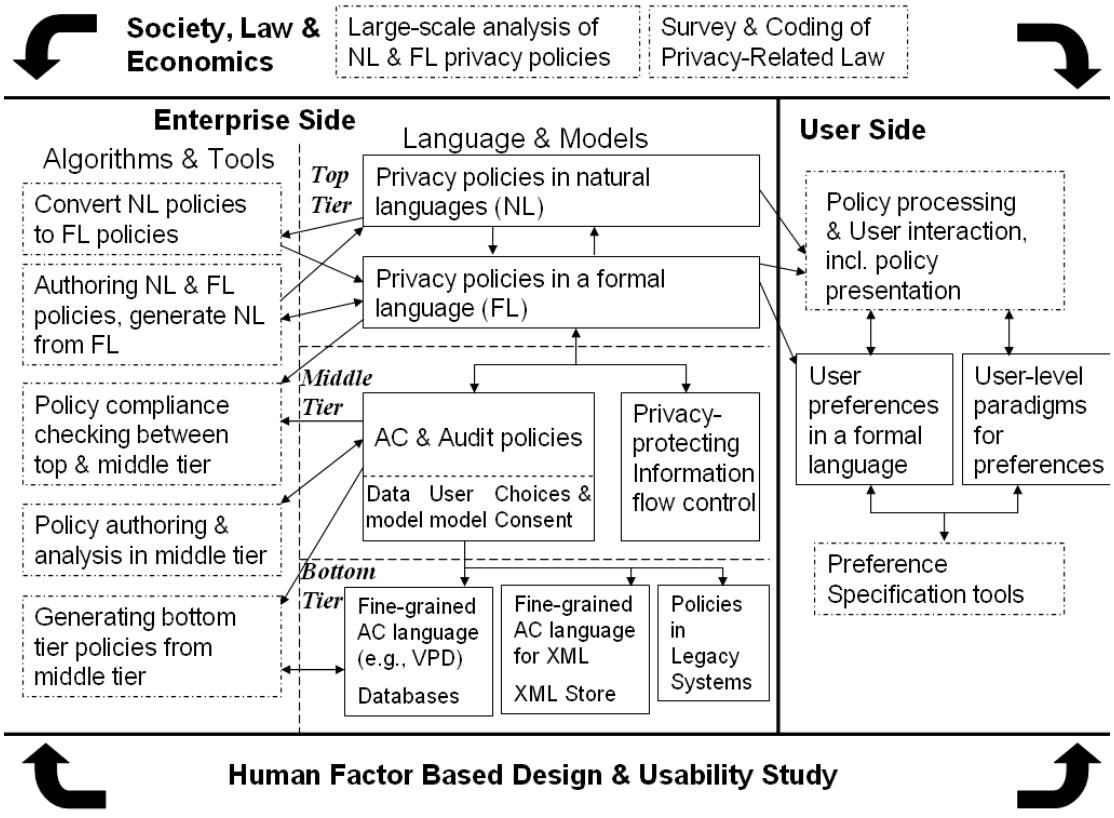
Society, Law & Economics
Large-scale analysis of NL & FL privacy policies
Survey & Coding of Privacy-Related Law
Enterprise Side
Language & Models
User Side
Algorithms & Tools
Top Tier
Privacy policies in natural languages (NL)
Privacy policies in a formal language (FL)
Policy processing & User interaction, incl. policy presentation
Convert NL policies to FL policies
Authoring NL & FL policies, generate NL from FL
Middle Tier
AC & Audit policies
Data model
User model
Choices & Consent
Privacy-protecting Information flow control
User preferences in a formal language
User-level paradigms for preferences
Policy compliance checking between top & middle tier
Policy authoring & analysis in middle tier
Bottom Tier
Fine-grained AC language (e.g., VPD)
Databases
Fine-grained AC language for XML
XML Store
Policies in Legacy Systems
Preference Specification tools
Generating bottom tier policies from middle tier
Human Factor Based Design & Usability Study

Figure 1: The architecture of a comprehensive framework for online privacy policy management.

Top tier (principles of privacy practices): An enterprise's high-level privacy promises are specified in privacy policies (using formal and/or natural language). Policies in this tier are intended for general Internet users. They should be specified by dedicated privacy officers who are familiar with both the enterprise's business practice and relevant privacy law and regulations. Key challenges include the design of a precise semantic model for privacy policies and expressive formal privacy policy languages. Policy languages for this tier should focus on which privacy goals are to be achieved, rather than how to achieve them.

Middle tier (security policies): In this layer, traditional security policies, e.g., those governing authentication, access control and information flow are needed to enforce high-level privacy policies. Policies at this tier should be specified by security officers who are familiar with high-level privacy policies and with the business processing needs of specific application domains. Within one application, privacy-centric access control and auditing policies ensure data access does not violate privacy policies or security requirements. Data models and user management models are needed to track how collected information is used by applications. A key challenge is the need to guarantee consistency between application-specific access control policies and privacy promises. Policy authoring and analysis tools based on specific application models are also required. Furthermore, because data may flow between applications that are governed by different high-level privacy policies, information flow control policies are needed to ensure that such data flow does not violate privacy promises.

The access control and auditing policies in this tier are application specific, but are usually independent of application implementation details. In an application, different levels of abstractions are commonly exploited to ease management overhead. For example, the model of information flow in an organization is usually independent of the physical storage of the information and the mechanisms through which information is exchanged between different departments. The separation of logical information flow and its physical storage and exchange implies the need for another level of privacy policy enforcement.

Bottom tier (enforcement in the physical layer): Access control and auditing policies need to be materialized through policy configurations in the underlying information repository. The nature of privacy policies tends to be fine-grained, e.g., each individual user may allow different usages of her data. Thus, fine-grained access control is needed; for example, if relational databases are used, then it may require row-level or even cell-level access control to support privacy constraints. An ambitious objective is to automatically generate fine-grained database access control and auditing policies from those in the middle tier, to eliminate potential logical errors during policy implementation. Furthermore, efficiency of policy evaluation and enforcement in the bottom tier is an important issue that needs to be addressed.

User Side: The user side components include user agents for preference specification and policy processing and presentation. The preference specification part interacts with the user through a paradigm that is close to the user's privacy protection

objectives and generates privacy preferences in a formal language, so that the matching between enterprises' privacy policies and users' preferences can be conducted automatically. The user agent also provides a more interactive user interaction model. When necessary, it presents the policies in an accurate and accessible manner and interacts with the user to help achieve privacy protection objectives.

Usability: This framework seeks to enable end-users to take an active role in protecting their privacy online; thus, usability is a key component. Because maintaining security and privacy is heavily reliant on users' cooperation (i.e., users need to specify their preferences), the maximal benefit of these preference specification methods cannot be realized unless interactions between the user and the system are simple and friendly. In particular, policy authoring and analysis tools as well as user agents need to be designed based on a comprehensive study of potential users' behaviors / preferences and existing tools.

Society, Law and Economics: Social norms and laws serve as the fundamental guidelines for enterprises to regulate their privacy practices and for users to establish necessary information disclosure principles. Although many organizations now post online privacy polices, these organizations must realize that simply posting a privacy policy on their website does not guarantee compliance with existing legislation. To date, privacy protection law in the U.S. includes coverage for healthcare data (the Health Information Portability and Accountability Act, HIPAA), information obtained from and/or about children (the Children's Online Privacy Protection Act, COPPA) and financial data (the Gramm-Leach-Bliley Act, GLBA). They not only regulate the collection and use of private information inside one organization, but also concerns about cross-organization information sharing.

If privacy regulations and laws are systematically analyzed and mapped to formal semantics, common privacy practice pitfalls can be avoided. Additionally, by studying users' social behaviors when accessing online services we will be better equipped to understand users' real privacy concerns — concerns which they do not articulate, but are evident in their behaviors. Societal studies benefit individual users and enterprises because it helps them design user-acceptable privacy policies. Finally, economic factors play important roles to promote the consideration of privacy and the adoption of privacy protection technologies, especially in the enterprise side. There is a need to study enterprises' and users' behavior from the economic perspective.

Research Issues

Specification of Privacy Policies

Privacy policies in the top tier are contracts between enterprises and end-users. A language for expressing such contracts must have an unambiguous semantics and significant expressive power. As discussed above, existing specification languages for privacy policies lack both. Relevant research issues that need to be addressed in this context include the following:

Development of a formal language for specifying privacy policies. Although P3P's limitations have been widely acknowledged [HJW03, Sch02, SHW02], the exact limitations have not been clearly identified and no comprehensive solution has been

proposed. A recent analysis of over 100 privacy policies in three different domains, e.g. general e-commerce, healthcare and financial websites [AEB04], has yielded over 1,000 goal statements and identified the goals appearing most frequently in textual policies. Most of these goals cannot be expressed in current privacy languages and thus they can be used to drive the development of more expressive formal privacy languages.

It is also critical to develop expressive privacy policy languages with an unambiguous semantics, serving as a semantic foundation for natural language privacy policies. An initial approach towards the definition of such semantics has been recently proposed [LYA03], based on which integrity constraints are introduced to maintain a P3P policy's semantic consistency. That approach focuses on providing a formal semantics for P3P, rather than remedying other weaknesses of P3P. It is however possible to build on that work in order to develop more expressive languages for specifying privacy policies, with a precise and clear relational semantics.

Automatic translation from natural language privacy policies to formal language policies. Although formal-language policies are being developed and deployed, it is unlikely that they will replace natural-language privacy policies in the foreseeable future. Using existing natural language processing software, tools can be developed to translate natural privacy policies into formal-language policies. Such tools would also facilitate the automatic generation of formal-language policies to and from natural-language policies, and consistency checking between formal and informal policies as well as within a natural-language policies, and would enable large-scale processing of online natural language privacy policies.

Enforcement and Auditing of Privacy Policies

To guarantee an enterprise's systems are in compliance with its privacy policies in the top tier, privacy constraints need to be integrated into specific applications in the middle tier, so they can be effectively enforced in business operations. An enterprise often provides several services to its users, and information flow frequently happens between different applications. Thus, privacy policy enforcement and auditing should be considered not only in the context of a single application but also in the context of the information exchange between different applications/systems. The recent JetBlue Airways privacy breach further motivates this requirement [AHB04].

Top tier privacy policies are abstract and thus cannot be directly enforced in the middle tier. Thus, we need to refine and materialize top-tier policies and map them into the relevant application domains. In particular, it is necessary to (1) specify middle-tier privacy policies based on specific application models; (2) verify their consistency with top-tier policies; and (3) integrate middle-tier privacy policies with access control policies of underlying data management systems which ultimately control private information access. Relevant research issues that need to be addressed in this context include the following:

Development of policy languages for specifying access control and auditing policies. Privacy protection requires either the design of new access control models or significant enhancement to current models. Most privacy policies allow users to decide whether to opt-out or opt-in to certain data usages; thus, a user's choices and consents have to be stored and used to make access control decisions. As a result, the access requirement

depends on both an enterprise's policies and user's choices. A language is needed to enable such highly fine-grained access control policies. The policy language should also specify auditing requirements for data access, so that an audit trail can be generated. One research problem is the selection of an abstract data model. Another research problem is the selection of a user model that allows access based on the attributes of users (e.g., the roles the user is playing, the tasks the user is current undertaking).

Theory and tools for comparing top-tier and middle-tier policies. An enterprise needs to ensure that middle-tier policies correctly enforce high-level policies. High-level privacy policies should not change with middle-tier policies. On the other hand, because middle-tier policies contain more information than high-level policies, they cannot be automatically generated from high-level policies either. It is likely that policies in the two tiers are specified independently; thus theory and tools for checking policy compliance need to be developed. Such tools will ensure that auditing policies in the middle tier are sufficient to generate an audit trail proving that they are in compliance with high-level policies.

Algorithms & tools to automate translation from middle-tier to bottom-tier policies. Efficient enforcement of middle-tier policies requires the use of native access control and auditing mechanisms provided by the data storage program (e.g., databases). The Virtual Private Databases (VPD) feature in Oracle provides fine-grained access control as well as auditing by dynamically executing a policy, which is a PL/SQL program, and attaching the generated predicate to each query. While this allows very flexible policies, authoring policies involves writing complicated procedure programs — a highly error-prone process. Furthermore, it is difficult to verify whether the policies are implemented correctly. Therefore, a mechanism is required to automatically translate middle-tier policies into physical repository policies.

Theory for information flow control based on privacy policies. Different enterprise sectors often have different privacy policies in place. Such heterogeneity comes from several sources. Global enterprises may be subject to privacy laws from different countries. Company mergers may result in enterprises with distributed and heterogeneous information systems, which in turn may have heterogeneous privacy policies. However, because the various sectors of an enterprise are often interconnected, the information flows among these sectors must be properly controlled to prevent privacy breaches [AHB04]. A key step in addressing this is the definition of a lattice based on privacy policies. This lattice definition will entail investigation of criteria and techniques for policy comparison. It is also important to investigate the extent to which the theory of information flow developed for MAC [BLP73] can be applied. To actually deploy information flow control techniques, one must properly define the interacting entities in an information flow process. Such interacting entities can be defined in various ways — according to organizational functions or a technical point of view (i.e., an entity can be an application program or a database system). Finally, a general notion of privacy contexts, which can be defined as a component within an organization characterized by a homogenous privacy policy with respect to a given sets of data is needed.

Privacy Management for the End-User

Privacy policies need to be communicated to end-users, enabling them to make meaningful decisions about whether to provide personal data online. However, just having the privacy policy in machine readable form is only a first step towards enabling end-users to control their privacy. We need to develop a user interaction model and a user agent that interacts with the user through high-level objectives. Relevant research issues that need to be addressed in this context include the following:

Development of a paradigm for specifying privacy preferences. This paradigm should be close to users' privacy objectives, rather than close to the data collection policies. Technical aspects of data collection and usage are often too complicated for users to fully comprehend. We conjecture that users' preferences should not be specified in terms of sharing specific data items, but rather in achieving privacy objectives. This paradigm should take into consideration users' limitations – it should be able to protect users from their own errors. The paradigm will account for privacy preferences that may vary for different transactions and websites. One possibility is to organize a set of preferences based on users' goals and websites' trust levels.

Methods and tools to present privacy policies to end-users in a uniform and accessible way. The P3P effort is predicated on the belief that privacy policies are too difficult for humans to understand; thus they are encoded in machine-readable form, which is then automatically processed by tools. We envision many cases in which human users would like to read the policy before entrusting their sensitive information to a website, rather than having a tool automatically make the decision for them. Instead of presenting users with pages of text that are laden with legal terms and not understandable to the majority of Internet users [AEB03], privacy policies should be presented in summary form for the users. Once the most significant axes of users' privacy concerns and goals are determined, we can determine how to best structure, organize and present this information to end-users. For example, the presentation may include scenarios of

what the company can do, warn possible negative consequences, or stress differences with existing preferences.

Privacy Policy: Legal and Economic Perspectives

Existing privacy policies are largely driven by organizations' legal concerns. Moreover, different organization's policies address different issues, despite being in the same industry [AEB04]. This suggests that companies within the same industry have different interpretations of the law or that errors of omission are common in privacy policies. In either case, while writing policies to address legal concerns is an understandable and prudent practice, it often leads to a mismatch between users' concerns and the information organizations disclose. Just as a law must survive constitutional challenge, a specified system should be demonstrably policy-compliant. Part of the solution to helping financial institutions become GLBA compliant is for organizations to be able to show that policies meet the requirements of the law, and that they are complete and unambiguous [AEB04].

Summary

Privacy is increasingly a major concern that prevents Internet users from fully enjoying the convenience, variety and flexibility offered by online services. A variety of privacy enhancing technologies has been proposed. While some technologies aim at preventing attacks that breach users' privacy, privacy policy technologies assume a cooperative relationship between service providers and users. Privacy policies allow enterprises and Internet users to communicate and negotiate privacy practices, and make online service privacy-aware. The proposed framework identifies key research challenges for the deployment and management of privacy policies. The framework shows that addressing these challenges will require close collaboration between academia and industrial researchers from multiple disciplines.

References

[AEB04] A.I. Antón, J.B. Earp, D. Bolchini, Q. He, C. Jensen and W. Stufflebeam. The Lack of Clarity in Financial Privacy Policies and the Need for Standardization. *IEEE Security & Privacy,* 2(2), pp. 36-45, 2004.

[AHB04] A.I. Antón, Q. He and D. Baumer. The Complexity Underlying JetBlue's Privacy Policy Violations. *IEEE Security & Privacy*, to Appear.

[AHK03] P. Ashley, S. Hada, G. Karjoth, C. Powers and M. Schunter. Enterprise Privacy Authorization Language (EPAL 1.1). *IBM Research Report*, October 1, 2003.

[BLP73] D. Bell and L. LaPadula. Secure Computer Systems: Mathematical Foundations. *Technical Report MTR-2547*, Vol. 1, MITRE Corporation, March 1973.

[Cran02] L. F. Cranor. *Web Privacy with P3P*. O'Reilly, 2002.

[HJW03] G. Hogben, T. Jackson and M. Wilikens. A Fully Compliant Research Implementation of the P3P Standard for Privacy Protection: Experiences and Recommendations. In *Proceedings of the 7th European Symposium on Research in Computer Security (ESORICS 2002)*, LNCS 2502, pages 104-125, Springer, October 2002.

[KSH03] G. Karjoth, M. Schunter and E. Van Herreweghe. Translating Privacy Practices into Privacy Promises - How to Promise What You Can Keep. In *Proceedings of the 4th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY 2003)*, pp. 135-146, June 2003.

[KSW02] G. Karjoth, M. Schunter and M. Waidner. Platform for Enterprise Privacy Practices: Privacy-Enabled Management of Customer Data. In *Proceedings of the Second International Workshop on Privacy Enhancing Technologies (PET 2002)*, LNCS 2482, pp. 69-84, 2003.

[LYA03] N. Li, T. Yu and A. I. Antón. A semantics-based approach to privacy languages. *CERIAS Technical Report TR 2003-28*, Purdue University, November 2003.

[Mar02] M. Marchiori (editor). The Platform for Privacy Preferences 1.0 (P3P1.0) Specification, W3C Recommendation, April 2002.

[P3P] W3C. Platform for Privacy Preferences (P3P) Project. http://www.w3.org/P3P/

[Sch02] D. M. Schutzer. Citigroup P3P position paper. Position paper for W3C Workshop on the Future of P3P. Available at http://www.w3.org/2002/p3p-ws/pp/ibm-zuerich.pdf

[SHW02] M. Schunter, E. Van Herreweghen and M. Waidner. Expressive Privacy Promises – How to Improve the Platform for Privacy Preferences (P3P). Position paper for W3C Workshop on the Future of P3P. Available at http://www.w3.org/2002/p3p-ws/pp/ibm-zuerich.pdf.