Advanced Metadata for Privacy-Aware Representation of Credentials

P. Ceravolo, E. Damiani, S. De Capitani di Vimercati, C. Fugazza, and P. Samarati
{ceravolo,damiani,decapita,fugazza,samarati}@dti.unimi.it

Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
via Bramante 65, Crema

Abstract

Semantic Web languages like OWL and RDFS promise to be viable means for representing metadata describing users and resources available over the Internet. Recently, interest has been raised on the use of such languages to represent individual data items contained in Personally Identifiable Information *(PII), supporting fine-grained release. To achieve this goal, the informative content of a credential must be dissected into atomic components so that users can selectively single out those to be released. In this paper, we outline methodologies for taking advantage of a distributed ontology-based framework for controlled release at both policy writing and evaluation time.*

1. Introduction

Nowadays the World Wide Web reaches the widest audience ever conceived through a broad range of devices such as computers, phones, and PDAs. Security and privacy concerns are increasingly important in this environment, where controlling the release, retention, and secondary use of personal data have become key issues. Advanced modeling of *Personally Identifiable Information* (PII) (i.e., any kind of information that can be linked to a specified individual) allows for controlling data release according to users' privacy requirements. On the other hand, such a model can assist system administrators in the specification of the information required for a resource or service to be granted. The structure referenced by policy requirements is rooted on a set of application-dependent elements referencing the formal definitions of credentials to be stored and requested by the system. This part of the knowledge base is aggregated from a pool of heterogeneous normative sources and constitutes the domain knowledge the negotiating parties are required to agree upon.

Privacy policies can then constrain the disclosure of PII and are associated with either instances of credentials (e.g., the VISA card) or abstractions defining them (credit card information). The latter allows multiple instances of the same credential to share the same preferences but are not required to. Moreover, data items within an instance of credential (e.g., the name on the credit card) or the corresponding fields in the underlying definition can be singled out to reach the finest granularity. Furthermore, correspondences can be drawn between the attributes of distinct credentials (e.g., the user's name) by mapping their definitions to a common structure. This way, privacy preferences can span along multiple credentials.

To refer and reason about credential we exploit the *base data schema* of the *Platform for Privacy Preferences* (P3P) [10], an XML-based standard language for expressing data-collection and data-use practices in a standard format. The P3P base data schema provides us with a well understood type-space for the definition of cross-cutting properties linking semantically equivalent data items. Unfortunately, P3P data schemata still lack the expressive power and the clearly defined semantics required for the definition of complex user credentials (a preliminary assessment of recent work on this issue is presented in [13]). Semantic Web languages like OWL [12] and RDFS [11] lend themselves very well to advanced representation of personal information inasmuch they allow for integrating credentials' structural definitions with a data schema expressing the meaning of the information to be exchanged, thus defining cross-cutting relationships linking semantically equivalent data items (e.g., birth dates) appearing in multiple credentials (e.g., a passport and a driver license).

In our previous work [4] we showed how the expressive power of standard XML-based Access Control languages can be increased to take advantage of ontology-based descriptions of the resources to be protected. Here, we address the problem of using ontology to model the portfolio, that is the entity enclosing all the sensitive data stored by the sys-

tem at both sides of the transaction. Specifically, we present some techniques allowing for the informative content of a user credential to be decomposed into atomic components, so that users can non-ambiguously single out items to be released.

2. Modeling the portfolio

With our model, we aim at integrating *declarations* (i.e., uncertified data provided by the user itself) and *credentials* (i.e., certificates signed by third parties) in the same context and specify preferences over them so that transactions can be carried out with the least recourse to the user intervention and limited disclosure of data [1]. Declarations represent personal data provided by the end user and stored by the digital identity management system for later use. Credentials are digital certificates signed by authorities notifying properties an user can provide to request a service access. Note that no assumption is made on the actual format of the credential, whether it is provided as a whole or it is possible to enucleate single elements (e.g., the `date-of-birth` out of a birth record). Furthermore, our model fully supports the use zero-knowledge proof technologies such as the Idemix [9, 2] credential system to reduce the need for the actual release of data.

Figure 1 depicts a fragment of a sample `Portfolio`, an entity enclosing all the sensitive data stored by the system. For the sake of simplicity, here classes correspond to credential definitions while attributes correspond to individual data items contained within, regardless of whether they represent more complex structures (e.g., the `expiry-date` made of day, month, and year) or else a ground data type. To help disentangling the different nature of entities, we capitalize class names (using CamelCase for composite names) and stick to the hyphenated syntax of P3P for attributes. We use uppercase names for actual instances of the defined data items, such as the `PDM05` attendance certificate. When not labeled, relations are of type *subclass-of*. We introduce this relation to avoid defining redundant attributes and to provide visual cues of dependencies: a policy specified on a class applies to all its subclasses and instances. For instance, with reference to Figure 1, any policy rule specified on structure `CreditCardInfo` will also apply to all of its descendants. Figure 1 shows different kinds of entities:

1. *Built-in entities expressing the system's functional requirements*, such as class `Profile` allowing the storage of information according to a given user profile so that multiple users can share the same data. For instance, a single credit card could be shared by a whole family, possibly with constraints on the amount that can be charged.
2. *Entities describing the inner structure of credentials and grouping declarations into classes* according to a shared ontology built from various trusted sources. For instance, class `CreditCardInfo` represents the standard information associated with a credit card.
3. *Entities representing the composition of atomic data items* and more general classes into higher-level abstractions, giving the user a way to categorize data in a custom fashion. For instance, class `AttendanceCertificate` is created to arbitrarily group a set of `AccreditationCertificate`s.
4. *Entities embodying instances of concepts* such as the actual values representing the user's `VISA` credit card.

In the example of Figure 1, it is clear that the personal data owner intervention was not limited to grouping data items or classes enveloping them with custom concepts such as `AttendanceCertificate`. She has also enriched the normative definition of class `CreditCardInfo` by sub-classing it with classes `BusinessCard` and `PrivateCard` and then extending the latter with the user-defined attribute `financial-agent`. Finally, she added to the single instance `MASTERCARD` the user defined attribute `customer-service`, pointing to a set of contact information. Correspondingly, policy definition languages should allow this kind of flexibility in the definition of data items to be requested or protected. By using a namespace-aware representation format the structures defined by the application, trusted parties, and the end user herself can be integrated with each other. We opted for OWL because its semantics provides a clean separation between classes and instances w.r.t. RDFS [8]. Classes will be primarily targeted by the access control rules defined by a system administrator for regulating the disclosure of resources. Instances will only be affected by privacy rules regulating the disclosure of sensitive information.By referencing the knowledge base described above, users can apply Privacy preferences to personal data items and service providers can define the requirements to be met by clients. Still the model lacks the facilities for expressing correspondences according to the semantics of data: the user could wish to protect a specific piece of information (e.g., her first name) without being forced to single out within the portfolio all the actual values bearing this information. To accomplish this, in our model we integrate a normative definition of credentials with an ontology-based representation of the P3P base data schema, obtaining an example of domain-independent categorization of personal information. By referencing an element from this structure, system administrators can automatically indicate a wide range of alternative credentials to be suggested to clients.

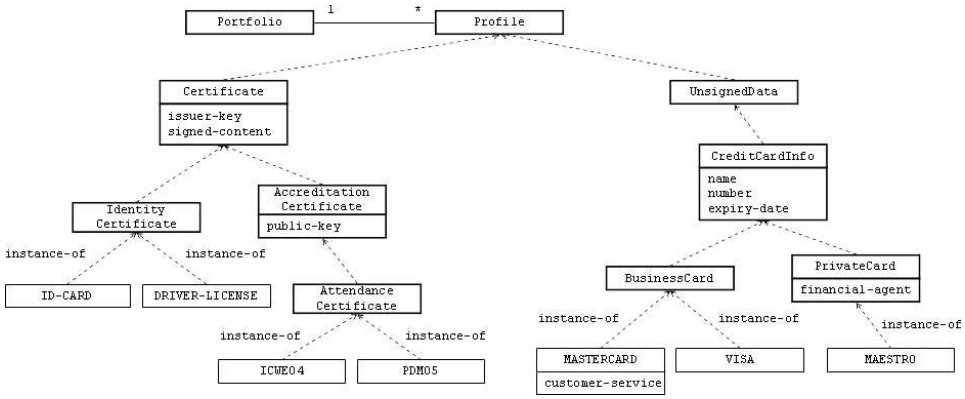

Figure 1. A sample portfolio.

3. The role of P3P

The core concept of P3P data schema is *data element* representing a single data item that can be either a root (unstructured) value or a more complex *data structure* composed of a set of data elements. An example of this is the definition of the `personname` data structure in the P3P text illustrated in Figure 2(a).

P3P data elements (and structures) are grouped into four *data element sets* (`user`, `thirdparty`, `business`, and `dynamic`). Note that each data element can appear in more than one structure. Figure 2 shows a portion of P3P base data schema definition clearly showing multiple references to the `personname` data structure. This can also be seen in the definitions of the `user`, `contact`, and `postal` data structures. The class diagram in Figure 2(b) shows *i)* the P3P data elements referencing the structure enclosing them (via a *part-of* relation) and *ii)* the structure defining them (via a *is-a* relation). In the figure, we use thick boxes for data structures (e.g., `contact`) and thin boxes for data elements (e.g., `name`).

3.1. An abstract model for P3P

We are now ready to provide a structural definition of the P3P semantics by arranging in a single structure the data element sets and the data structures composing them (such as `date`, `login`, and `http-info`), down to the single data element. For defining the architecture we faced two possible options: *i)* expressing the portfolio with the OWL representation of P3P data schemas; *ii)* expressing the portfolio with a custom representation format and linking them with the P3P base data schema. In the first case data elements in the base data schema (e.g., the properties associated with `Personname`) will be instantiated to represent actual values enclosed in credentials. In the second case we associate elements of the P3P base data schema with portfolio's instances and definitions. We opted for the second choice because we need to link actual data values with their meaning but wish to avoid the semantic and structural constraints of P3P. As shown in Figure 1, the informational content of users' credentials is modeled in a similar way, taking advantage of *is-a* sub-typing to represent variations of a base credential. For instance, legislation of different countries may require different data elements to appear within the same credential.

4. Using semantic web languages for representing heterogeneous personal information

Our model of the P3P data schema can be used as a base ontology for expressing the meaning of data contained in the user's portfolio. Representing the P3P model according to an OWL syntax [12] we enrich our model with an interpretation that dramatically reduces ambiguity. For instance, the OWL interpretation can distinguish between *part-of* and *is-a* relations, taking advantage of the reasoning features associated with the language (e.g., as shown in [3, 7]). Also an OWL reasoner allow to set custom rules where the user can define equivalences bounded to his applications. Using Portfolio to support policy language evaluation, we can extend a policy rule including a subject description involving a data element to descriptions in terms of other data elements sharing the same semantics. For instance, consider the following portion of an XACML subject descriptor.

```
<SubjectMatch
  MatchId="urn:oasis:...:xacml:1.0:function:string-equal">
  <AttributeValue
    DataType="http://www.w3.org/2001/XMLSchema#string">
    USA
  </AttributeValue>
  <SubjectAttributeDesignator
    AttributeId="urn:ourdomain:attribute:Postal.country"
    DataType="http://www.w3.org/2001/XMLSchema#string"/>
</SubjectMatch>
```

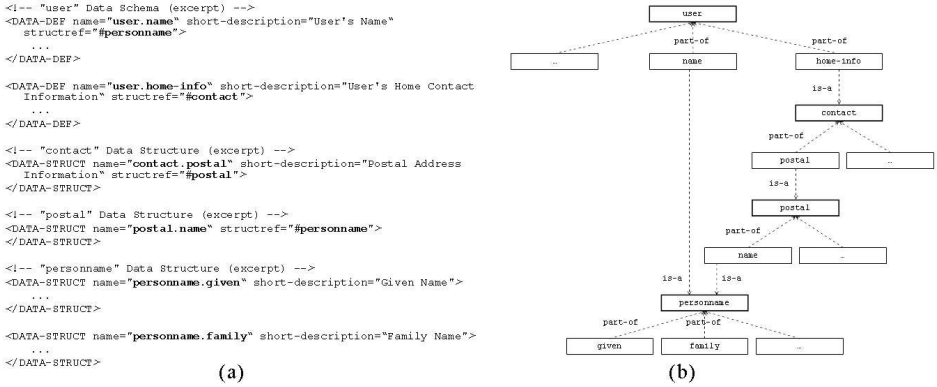

Figure 2. A small section of the P3P base data schema: (a) XML description, (b) graphical representation.

Here, the attribute being referenced does not represent by itself any credential envisaged in the Portfolio. Nevertheless it can be mapped to the attributes bearing the same information in the identity card, driver license, and passport. This way a single attribute can unfold a whole set of alternatives. Furthermore, it is possible to enrich our ontology with a specific task ontology representing a credential as a whole, relating it to the enclosed data elements, and adding facilities for the diachronic evolution of its normative definition (e.g., the migration process of banking records due to the Euro's introduction).

5. Conclusions and future work

This paper has outlined a structural model of P3P data schema semantics and illustrated its encoding in terms of Semantic Web languages like OWL. It represents only a first step towards a semantics aware access control, and much work is still to be done before this encoding can be used in practice. A necessary improvement is mapping the policy preference language to the OWL syntax so that policies and requirements associated with them can be exchanged as triples without translating policies from the original format to the corresponding OWL representation.

Acknowledgments

This work was supported in part by the European Union within the PRIME Project in the FP6/IST Programme under contract IST-2002-507591 and by the Italian MIUR within the KIWI and MAPS projects.

References

[1] P. A. Bonatti, P. Samarati - *A Uniform Framework for Regulating Service Access and Information Release on the Web* - Journal of Computer Security 10(3): 241-272 (2002).

[2] J. Camenisch, and E. Van Herreweghen - *Design and Implementation of the idemix Anonymous Credential System* - IBM Zurich Research Laboratory - http://www.zurich.ibm.com/ jca/papers/camvan02.pdf

[3] L. Cranor, B. McBride, and R. Wenning - *An RDF Schema for P3P* - World Wide Web Consortium, Note. 25 January 2002 - http://www.w3.org/TR/p3p-rdfschema/

[4] E. Damiani, S. De Capitani di Vimercati, C. Fugazza, and P. Samarati - *Extending Policy Languages to the Semantic Web* -In Proc. of the International Conference on Web Engineering, Munich, Germany, July 2004.

[5] E. Damiani, S. De Capitani di Vimercati, P. Samarati - *Managing Multiple and Dependable Identities* - IEEE Internet Computing 7(6): 29-37 (2003).

[6] *eXtensible Access Control Markup Language* (XACML) - Organization for the Advancement of Structured Information Standards - http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

[7] G. Hogben - *P3P Using the Semantic Web* - World Wide Web Consortium, Note. 16 August 2004.

[8] I. Horrocks and P.F. Patel-Schneider - *Three theses of representation in the semantic web* - In Proc. of the Twelfth International World Wide Web Conference (WWW 2003), Budapest, Hungary, May 2003.

[9] *Idemix anonymous credential system* - IBM Zurich Research Laboratory - http://www.zurich.ibm.com/security/idemix/

[10] *Platform for Privacy Preferences* (P3P) - World Wide Web Consourtium - http://w3.org/P3P/

[11] *RDF Vocabulary Description Language* (RDFS) - World Wide Web Consourtium - http://www.w3.org/TR/rdf-schema/

[12] *Web Ontology Language* (OWL) - World Wide Web Consourtium - http://w3.org/2004/OWL/

[13] T.G. Yu, N. Li, A. Anton - *A Formal Semantics for P3P* - In Proc. of the ACM Workshop on Secure Web Services, Washington, USA, October 2004.