AN ONTOLOGICAL APPROACH TO VERIFYING P3P POLICIES*

Assadarat Khurat1 and Boontawee Suntisrivaraporn2
1Institute for Security in Distributed Applications, Hamburg University of Technology, Hamburg, Germany
2School of Information and Computer Technology, SIIT, Thammasat University, PathumThani, Thailand

Keywords: P3P policy, Ontology, Semantic web.

Abstract: Privacy has become a crucial issue in the online services realm. P3P policy is a privacy policy enabling websites to express their privacy practices. With this policy, online users can check against their privacy preferences which facilitates the users to decide whether or not the service should be used. However, the interpretation of a P3P policy is unwieldy due to the lack of a precise semantics of its descriptions and constraints. For instance, it is admissible to have purpose and recipient values that have inconsistent meaning. Thus, there is a need for an explicit formal semantics for P3P policy to mitigate this problem. In this paper, we propose to use an OWL ontology to systematically and precisely describe the structures and constraints inherent in the P3P specification. Additional constraints are also defined and incorporated into the ontology in such a way that the reasons of an invalid P3P policy can be disclosed after the verification done by an OWL reasoner.

1 INTRODUCTION

Privacy has become an important issue for the online world. To provide a service, online service providers may collect and store users’ sensitive data where misuses of these data cause privacy breaches. Many countries and organizations, thus, have concerned with privacy issue seen from enactment of privacy laws—e.g. Privacy Acts in the USA, EU Directives in European Community and OECD Guidelines for international level.

The Platform for Privacy Preferences (P3P) Policy (Cranor et al., 2002), standardized by W3C, is a technology that stems from this privacy concern. It can be used by websites to express their practices about customers’ data in the machine-readable format, XML. A P3P user agent embedded in e.g. a web browser can compare P3P policies of service providers with the users’ privacy preferences specified beforehand. The comparison result enables the users to decide whether to use the services or not. However P3P policies may contain internal semantic inconsistencies. Thus, to detect existing discrepancies and regain consistency, the formal semantics for P3P is compulsory and it needs to be explicitly formalized.

The Web Ontology Language (OWL) (Bechhofer et al., 2004), a W3C recommendation, is a well-known semantic web technology. Due to its capability in expressing logical formalism (Description Logic); and both structures of P3P policy documents and dependencies that can be described as an ontology, we decide to use OWL ontology to provide formal semantics for P3P. The benefits of employing OWL for P3P are twofold: (i) the logical underpinning of OWL guarantees preciseness of the definitions and constraints, i.e. ambiguity is reduced; and (ii) an OWL reasoning tool can be exploited to automatically check consistency of a particular P3P policy. Our proposed framework is based on the data–purpose centric interpretation. We also aim to be able to detect inconsistencies in a P3P policy, and to explain which part is the culprit.

2 P3P & ITS POTENTIAL INCONSISTENCIES

In P3P policy, not only how websites treat the collected data is expressed, but other aspects concerning privacy practices can be also described. These aspects are Entity, the policy issuer; Access, the ability of individuals to access their data; and Dispute-Group, resolution procedures when disputes between privacy policies occur.

How the websites may deal with the collected data
is described in Statement which is the problematic part inspiring this work. A policy can contain one or more Statement elements where each Statement consists of Data-Group, Purpose, Recipient and Retention. The Data-Group element contains a list of data (Data element) which the services may collect and optionally data categories (Categories element). P3P specifies the categories for its defined standard set of the Data elements. The data standard set is structured in a hierarchy and grouped in four sets; dynamic, user, thirdparty and business. Some Data elements can be placed in more than one group. The elements Purpose, Recipient and Retention describe, respectively, for which purpose the data may be used, to whom the data may be distributed, and for how long the data will be kept. The Purpose and Recipient elements can have multiple values while the Retention element can have only one value. P3P specification defines twelve values for Purpose, six values for Recipient and five values for Retention.

Besides the main elements, Web sites/services can inform their users which data element, which purpose of data usage, and which data recipient are either optional or mandatory through an optional attribute called Optional (yes or no) for the former and Required (always, opt-out or opt-in) for the latter two.

An example P3P policy of walmart.com, consisting of two statements (S1 and S2) is shown in Fig.1. S1 collects user’s contact information and allows her to create an account. S2 collects other personal information, viz. name, email, postal address for conducting line interaction”. This introduces a conflict since the data collected under purpose develop are required to be stored for longer than permitted time no-retention.

3 DATA–PURPOSE CENTRIC SEMANTICS FOR P3P

In order to establish an Ontology, the relationships between entities in the domain must be known. In P3P policies, it is certain that the Data element is a main entity. The work from Ting Yu et al. (Yu et al., 2004) proposed a formal semantics for P3P employing a data-centric view. However, the purpose of data usage is also an important information for data practices, i.e. there must be a reason to collect the data. In addition, how long the data should be retained depends on the purpose of collection. Moreover, this way of interpretation also complies with the Purpose Specification Principle of OECD and the EU Directive 95/46/EC Article 10(b) that requires the data controller (website) to inform the data subject (user) at least about the identity of the controller and the purposes of the data collection. We, therefore, propose to use both the data and purpose as the keys in our formal semantics for P3P.

Besides the inherent constraints according to P3P specification, we define additional constraints for checking potential semantic conflicts described in previous section as follows:

Multiple Statements. The elements that should have only one value are Retention element; and Op-
tional and Required attributes. Under the data–purpose based interpretation, we define that in a policy there must be only one value of Retention and Required for each data–purpose pair, otherwise the policy is considered invalid. The constraint for Optional attribute is defined analogously but only for each data, since this attribute only belongs to the Data element.

Data Hierarchy. Considering data standard set’s hierarchy, it does not make sense if the data has more restrictions on its collection than its descendant. Therefore, we define that in a policy containing data where one (e.g. #user.bdate.ymd.year) is a descendant of the other (e.g. #user.bdate), the Optional value of the descendant must be equal or more restrictive than the other one; where we define that the value no is more restrictive than yes. The same condition also applies to the Required values of Purpose and Recipient elements for their constraints, where we define that the value always is more restrictive than opt-out, and opt-out is more restrictive than opt-in.

Optional Attributes. Due to unclear meanings of optional attributes (Optional, and Required) in the P3P specification, we define that, for each data, if all of its purposes are optional (Required value of Purpose element is opt-in), its collection must be optional (Optional value is yes). This is because, for opt-in, the services may use the data only when the users specifically request to. Thus, before this request is made, the services should not collect the data.

Inconsistent Meaning between Purpose, Recipient, Retention and Data Category Values. Except the pair between Data Category and Retention, we define eight constraints to check semantic consistency of each pair between Purpose, Recipient, Retention and Data Category. Four constraints are defined for the pair Purpose and Data Category according to the User Agent Guidelines (Cranor, 2003) which has been appended to P3P1.1 specification. For the rest pairs i.e. between Purpose and Recipient; Purpose and Retention; Retention and Recipient; and Recipient and Data Category, one constraint is defined for each. Due to space limitation we give an example of a constraint between Purpose and Retention as below:

In a policy, when Purpose value is one of admin, historical, develop, pseudo-analysis, pseudo-decision, individual-analysis, individual-decision, telemarketing and contact, its associated Retention value must not be no-retention.

4 AN ONTOLOGY FOR P3P

We propose to use an OWL ontology to systematically and precisely describe the structures and constraints inherent in the P3P specification. Once an ontology has been deployed, any P3P policy can be verified against this ontology with the help of an OWL reasoner. Our aim is to be able to verify whether a given policy is valid, and if not, what is wrong.

As shown in Sec. 2, a policy consists at least one Statement, which in turn comprises several elements e.g. Data, Purpose, Recipient, and Retention. Note that we focus only on these four elements for clarity of discussion.

An obvious modeling choice is to define a class for each of these elements and relate them with appropriate properties/roles. To make sure that the purpose for one data is not grouped with another data, we propose here to flatten original P3P statements such that each resulting reified statement has exactly one Data and one Purpose. The class Data represents any data item per se, whereas an additional class (CollectedData) represents those data collected by a policy for some purposes. Due to our proposed data–purpose centric model where the purpose of the collected data is considered important for data practices, we also define another class (DataPurpose-CollectionPractice) to represent the purposes for which the data are collected, as shown in Fig. 2. The corresponding OWL definitions of this model are given by \( \alpha_1 \)–\( \alpha_3 \) in Fig.3. At the bottom of Fig.3 are role axioms required for reasoning. The role inclusion axioms \( \rho_1 \)–\( \rho_5 \) and \( \rho_7 \) specify, respectively, that hasPart is a superrole of every other role and that it is transitive. The hierarchical structures of data in P3P are organized using an aggregation role hasSubDataStructure, and every leaf data item relates to their corresponding data category via another role categorizedBy. This design enhances the modeling in (Damiani et al., 2004; Hogben, 2005) by adding the left-identity role inclusion axiom \( \rho_6 \). In the presence of this axiom, any category of a sub-data
structure is automatically propagated to its super-data structure.

In general, constraints shown in the previous section can be translated into a logical expression which then form (part of) a definition in the ontology. However, checking constraint violations in any given P3P policy by this approach is insufficient to explain what is wrong in the policy. We thus propose to define classes (called InvalidPolicy) with specific definitions to represent these constraint violations, instead of specifying logical expressions directly in the ontology. This modeling decision enables us not only to detect the policy invalidity but also to know the underlying reasons. We define twelve InvalidPolicy classes but, due to space limitation, only one is depicted here as $\beta_1$ in Fig.3. InvalidPolicy1 represents the class of invalid policies that have multiple retention values for the same data–purpose collection practice. Multiple retention values are captured with the help of at-least number restrictions. Since the data #user.login and #user.home-info of the policy in Fig.1 have two retention values, when we run an OWL reasoner (Hermit 1.3.3 in Prot´eg´e), the policy is inferred as a member of class InvalidPolicy1.

5 RELATED WORK

A work on formalizing P3P in an ontology (Hogben, 2004) was proposed as a W3C working group note. This and our work share the ideas of modeling most P3P entities as concepts (classes of individuals), of flattening P3P statements, of modeling data nested structures by an aggregation role instead of the subclass relation, and modeling data categories as superclasses. The modeling choice of this work differs to ours that each policy statement is flattened to a few refified statement objects where each describes a collection practice of a data item. Another subtle difference however remains in the choice between OWL quantifications. We reckon that a sensible policy should describe at least one collection practice of a data item, so some is chosen instead of only. In addition, we use roles subDataStructureOf and hasSubDataStructure in place of may-include-members-of, which is rather confusing. The fact that a super-data structure may or may not include a sub-data structure is modeled in our ontology using a number restriction.

Damiani et al. (Damiani et al., 2004) and Hogben (Hogben, 2005) proposed a way to represent P3P-based data schema in the Semantic Web, focusing on data schema of P3P 1.0. In these works, data items are similarly modeled as classes, but they are interrelated via three roles, viz. is-a, part-of, and member-of which is unnecessarily complex and error-prone.

6 CONCLUSIONS

We proposed an ontology model for P3P based on data–purpose centric view. Several constraints required to prevent certain semantic inconsistencies have been identified and formalized in an OWL ontology. Our constraint violation detection are implemented, instead of logical constraint, in such a way that can capture constraint in OWL classes which can provide reasons of P3P policy invalidity.

REFERENCES


