A Privacy Focused Formal Model of Authorization for Data Modeled using Semantic Web Technologies

Jenni Reuben
and Simone Fischer-Hübner

Karlstad University
651 88 Karlstad, Sweden
Email: [firstname.lastname]@kau.se

***Abstract*—Origin of digital artifacts is asserted by digital provenance information. Provenance information is queried for proof statement validations, failure analysis, as well as replication and attribution validations. The history of a data instance that specifies dependency among different data items that produce the data instance is better captured using semantic web technologies. However, such provenance information contains sensitive information such as personally identifiable information. Further, in the context of Semantic Web knowledge representation, the interrelationships among different provenance elements imply additional knowledge. In this paper, we propose an authorization model that enforces the purpose limitation principle (an essential data protection principle) for such semantically related information. We present the formalization of the security policy, however the policy does not directly conforms to the desired authorization outcome. Therefore, security properties for important relationships such as subset, set union and set intersection are defined in order to ensure the consistency of the security policy. Finally, a use case scenario demonstrating the defined security policy and the properties is presented to indicate the applicability of the proposed model.**

***Index Terms*—Semantic Web; Access control; Security; Privacy; Purpose binding; OWL; RDF.**

I. Introduction

Provenance is a well known concept in the art world, it refers to the documented history of an art object, which is used to evaluate the significance of the art object in relation to other similar objects [1]. Similarly, digital provenance describes how a digital object has been brought to its current state. Such provenance information accounts for proof statement validations, failure analysis, as well as replication and attribution validations. In support of various provenance related queries, access to the provenance traces is desirable. We consider that this functionality of an application is facilitated by a repository (or several repositories in the case of distributed environment) that stores provenance traces. Subsequently, regulated access to these stores has become a crucial requirement to prevent provenance data misuse.

Nevertheless, controls to enforce legitimate access to provenance information should account for two important factors of digital provenance. First, emerging applications of provenance such as semantic web, e-science and cyberinfrastructure [2] demands provenance information to be available on the web, interpretable by machines, effectively discovered and interoperable. This is evident from an array of PROV specifications - a recent standardization efforts from W3C Provenance Working Group, which are based on semantic web principles. Second, provenance information often contains sensitive information such as *i*)personally identifiable information [3]–[6] and *ii*) semantically revealing relationships of different provenance elements, which would enable inference of additional personal data [7]. Therefore, a comprehensive approach to the enforcement of access restrictions is required in which, both the privacy requirements of the stored data and the semantic richness of the involved data model are taken into account.

Most often than not, research in provenance access control tends to focus on identity-based authorization [8]–[10]. Little attention has been paid to rule-based authorization models, i.e., the models that take into account certain attributes of the data, which may be denoted in terms of security labels, usage purposes, etc. Furthermore, few research [11]–[14] have investigated solutions to the access regulation problem when the data is enriched with formal semantics.

The aim of the paper is to provide an authorization model that takes into account both the data privacy attributes and the semantic richness of the data model. The present work extends the previous formal privacy model by Fischer-Hübner [15] that enforces data protection principles such as *purpose binding* and *necessity of data processing*. In particular, in the current model, the degree of access restriction granularity is enhanced by introducing purpose hierarchy and additional security properties are defined for the semantically enriched data.

Specific contributions of the paper are as follows:

- We identify additional necessary components for enforcing authorization that is based on the purpose limitation principle. Further, we formalized the security policy that constrain the data access based on the purposes for which the authorization objects are collected (Section III). In particular, the consistency of the security policy in the presence of web ontology's class interrelationships are ensured by the security properties defined for subclass, class union and class intersection relationships (Section III-B).

 ISBN: 978-1-61208-558-6

- A use case is presented that demonstrates the applicability of the defined security policies and properties of the proposed model (Section IV).

This paper is organized as follows. In Section II, we describe the background knowledge of a Semantic Web information system and a brief introduction to the Task-Based Privacy model. Definitions of the required components for formalizing the authorization model, and for the formalization of the model's security properties are presented in Section III. Section IV demonstrates the applicability of the model by means of a use case scenario. The current state-of-the-art is analyzed in Section V followed by the conclusions of the paper, which is presented in Section VI.

II. PRELIMINARIES

In this section, we present the background knowledge on the Semantic Web query answering. Semantic web is a web that is targeted for automated reasoning, integration and interoperability. This is realized by enabling the machines to understand the information content. Nevertheless, the technologies that support Semantic web vision need to be weaved into the well-established web standards such as Universal Resource Identifier (URI) and eXtensible Markup Language (XML).

The starting point for making the machines understand the web contents is to give the contents a well-defined meaning [16]. Intuitively, knowledge representation technology from artificial-intelligence research provides an excellent way to define and to reason about things that exists in a domain of interest. Accordingly, the information in the web documents can be described, thereby providing a meaningful structure to the web contents. Given the meanings and the sets of inference rules, the machines can conduct automated reasoning. On a related note, this notion of adding structured and semantical annotations to actual data lends itself to the concept of provenance, which is a metadata describing data.

A. Semantic Annotations

The Resource Description Framework (RDF) specifies that the descriptions that annotate the web information take the form of a triple. The RDF triple form is similar to the subject, verb, object structure of an elementary English sentence [17]. Intuitively, this makes the descriptions to be readily encoded using XML tags [16]. Formally, the description asserts that a particular thing (some entities in the domain of interest) has a property (relation) with certain values (again referring to the entities in the domain of interest). A set of RDF triples is known as RDF graph and a collection of organized RDF graphs is called a RDF dataset. The RDF triple for the statement "John is a person" is shown in Figure 1. Correspondingly, Figure 2 shows the RDF graph, which represents the entities in the RDF triple as nodes and the relations as directed edges.

B. Defining the Semantics

As mentioned earlier in this section, meanings of the terms used in the semantic annotations are provided by another Semantic Web component called ontologies. Different from

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf
      -syntax-ns#">
  <rdf:Description rdf:about="http://example.
      com/~jwille#john">
  <rdf:type
    rdf:resource="http://example.com/person-
       voc#Person"/>
</rdf:Description>
</rdf:RDF>
```

Figure 1. RDF triple

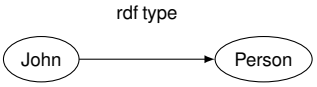

Figure 2. RDF graph

the ontologies in the metaphysical context, an ontology in the Semantic Web context refers to a computational artifact that formally defines and categorizes entities that exist in a domain of interest as well as explicate the relations between the entities. As a result, intelligent agents in the Semantic Web context can unambiguously deduce the meaning of the things in the world (domain of interest). The World Wide Web Consortium (W3C) Web Ontology Working Group standardizes the Web Ontology Language (OWL) as a formal language for representing ontologies in the Semantic Web. An example of parts of PROV ontology (PROV O) [18], which is encoded using RDF/XML syntax specifications is shown in figure 3.

Furthermore, as OWL ontologies are grounded on the formal logic using OWL 2 $DirectSemantics$ [19], additional inferences can be derived from the explicit declarations. For example, an ontology that includes the information that Mary is a mother and every mother is a women, implicitly specifies that Mary is a women.

Ontologies and ontology-based semantic annotations are used in many application scenarios such as Semantic Web search engines, provenance, etc. Access to the triples subsequently to the RDF graphs is facilitated by The Simple Protocol and RDF Query Language (SPARQL). SPARQL 1.1 [20] includes an extension point, which specifies OWL-based semantics for query evaluation.

However, this notion of semantic interpretations and derivation of implicit information introduce novel access control challenges. Access control mechanisms developed for XML do not readily lend themselves to the notion of automated reasoning of the RDF graphs. In the Semantic Web context, the authorization objects encompass relationships among entities and the additional RDF graphs that follow from such relationships rather than the structure of a web document. The model we propose is a mandatory access control whereby the active agents of the system must comply to certain rules for a successful access. The access policy of our model is based on the purpose of the access. It is evolved from the Task-Based Privacy model [15], which technically enforces essential data protection principles such as "purpose binding" and "necessity

```
<!DOCTYPE rdf:RDF[
  <!Entity owl "http://www.w3c.org/2002/07/owl#
      ">]>
<rdf:RDF xmlns:owl ="http://www.w3.org
    /2002/07/owl#"
  xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-
      syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-
      schema#">
<owl:Ontology rdf:about="">
  <rdfs:label>Example Ontology</rdfs:label>
  <rdfs:comment>An example ontology</
      rdfs:comment>
</owl:Ontology>
<owl:Class rdf:ID="Entity" />
<owl:Class rdf:ID="Agent" />
<owl:ObjectProperty rdf:ID="wasAttributedTo" /
    >
<owl:DatatypeProperty rdf:ID="value" />
<owl:Class rdf:ID="Person">
  <rdfs:subClassOf rdf:resource="#Agent">
</owl:Class>
</rdf:RDF>
<owl:Class rdf:ID="plan">
  <rdfs:subClassOf rdf:resource="#Entity">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#
          wasAttributedTo"/>
      <owl:someValuesFrom rdf:resource="&Pers;
          person"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Figure 3. An example of an OWL ontology encoding parts of the PROV data model

of data processing" .

C. Task-Based Privacy Model

The key idea of the Task-Based Privacy model is to place technical controls for enforcing data protection principles such as purpose binding and necessity of data processing.

The authorization policy is expressed in terms of the purposes that are assigned to the active agents and to the passive objects of a system. Data objects are categorized into object-classes, which are assigned a set of specified purposes representing the usage purposes of the data objects in those classes. The tasks that operate on those objects are also assigned appropriate purposes. Tasks are specific to an application, hence depending on the applications a task could comprises of several functionalities. Each task or simply a functionality serves exactly one purpose. The active agents are authorized to perform a set of tasks, which is further restricted by the kind of data transformations that are allowed for each task. Given these components, the central security property of the model is defined as:

Purpose binding property: An active agent is permitted to access a data object, only if the purpose of the task that the agent currently performs is contained in the set of purposes specified for the object-class that encloses the data object(s).

III. Privacy Preserving Authorization Model for Semantic Annotations

We consider, in this work that the data usage is constrained by the purpose limitation principle. This is required for complying with the European Union (EU) General Data Protection Regulation (GDPR) and other privacy laws. EU GDPR (Art.5 1(b)) mandates that the processing of personal data should only be permitted if it is necessary to serve the purposes for which the data is collected [21].

A. Model Components

Definition 1: (**Subjects S**) *A subject is an active agent of a system, which is properly identified and authenticated. S is set of active subjects.*

$$S = \{s_1, s_2, ...s_n\}$$

OWL ontology is a formal specification that includes categories of entities (referred to as classes in OWL terminology), and structured vocabularies that explicate the relationships among the classes. Relationships include both the taxonomy of classes and other relationships among the classes. A relationship other than the class taxonomy is the relation between the instance of two class, which are the domain and the range of the relationship.

Definition 2: (**OWL Ontology**) *An OWL ontology O is defined as a tuple,*

$$O = (C, T_c, R)$$

Where, C is a set of OWL classes $\{c_1, c_2, ..., c_i\}$, T_c is the set of terms that explicate the taxonomy of classes and R refers to the set of other types of relations among the instances I of the classes in C. $R = \{r_1, r_2 ..., r_n\}$, where r_i is a relation from c_i to c_j $(i \neq j)$, given that $Dom(r_i) = \{I_{c_i} \in c_i \mid \exists I_{c_j} \in c_j, (I_{c_i}, I_{c_j}) \in r_i\}$ and $Ran(r_i) = \{I_{c_j} \in c_j \mid \exists I_{c_i} \in c_i, (I_{c_i}, I_{c_j}) \in r_i\}$.

Definition 3: (**RDF triple**) *A RDF triple, which is of the form (statement-subject ($\mathcal{S}_{rdf}$), predicate ($\mathcal{P}_{rdf}$), statement-object ($\mathcal{O}_{rdf}$)) is an element of the* Cartesian product *of $((C \cup B) \times (T_c \cup R) \times (C \cup L))$.* Where,

- $\mathcal{S}_{rdf} \in (C \cup B)$, such that $\exists c_i \in C : \mathcal{S}_{rdf} \in c_i$ and B is a set of blank nodes.
- $\mathcal{P}_{rdf} \in (T_c \cup R)$.
- $\mathcal{O}_{rdf} \in (C \cup L)$, such that $\exists c_i \in C : \mathcal{O}_{rdf} \in c_i$ and L is a set of literals.

A RDF graph G is a finite set of RDF triples. However, in addition to the direct mappings of RDF instances to OWL classes and respective relationships, additional RDF triples and graphs are entailed. This is due to the interpretation of how the OWL classes and terms are connected in the direct-semantics-based OWL ontology. We refer to these additional RDF instances as entailments. Accordingly, the authorization objects is defined as follow;

Definition 4: (**Authorization objects**) *An authorization object is a subgraph of G that explicitly and implicitly conforms to the OWL ontology O under the OWL 2 direct semantics. A set of authorization objects is denoted as $\mathcal{O}_A$.*

An application is made up of several tasks. Examples tasks of a hospital information system are admission, diagnosing, surgery, care transfer, discharge, and billing. Definitions 5-10 are derived from the Task-Based Privacy model [15].

Definition: 5: (**Tasks**) *Tasks are operations through which the subjects access the authorization objects. T is a set of all tasks that are defined for a system.*

$$T = \{t_1, t_2, ...t_n\}$$

Each subject will be authorized for a subset of tasks either depending on their roles in an organization or on other contextual attributes.

Definition 6: (**Authorized tasks**) *A set of tasks that a subject is authorized to perform is provided by a task assigning function.*

$$AT : S \rightarrow 2^T \setminus \emptyset$$

where, $AT(s_i)$ is the set of authorized tasks of s_i.

We further distinguish the current task that the subject is performing from its authorized set of tasks. If there is no current task for a subject, then a standard value "Nil" is assigned as its current task.

Definition: 7: (**Current task**) *Current task is the task that a subject is currently performing,*

$$CT : S \rightarrow T \cup \{Nil\}$$

Information, specifically personally identifiable information, is collected and stored for certain usage purposes. As pointed out earlier, it is required by data protection laws (e.g., the EU GDPR) that the processing of such information should be permitted where it is necessary to serve those purposes. Accordingly, the tasks that access the information for processing must be assigned a specific purpose that they are designed to serve. As an example, in a hospital information system, an admission task, which is designed to serve the admission purpose is only allowed to operate on information that is collected for the administrative purpose.

Definition: 8: (**Purposes**) *The set of all purposes for which data is collected and processed in an application is denoted by P*

$$P = \{p_1, p_2, ..., p_n\}$$

Relationship between the purposes play a crucial role. Hierarchically structured purposes can, *i*) improve the granularity of the access control rules by constraining the access to specific sub purposes and *ii*) lend itself well to specify interconnected purposes to related OWL classes. Set P is defined to have an order relation $\leq$ and forms a partially ordered set (P,$\leq$). The hierarchical structure we propose for the purposes is exemplified and illustrated in figure 4. The top level nodes are "super" purposes that dominates their children. As a consequence, tasks and authorization objects are identified both from the purposes directly assigned to them and from the purpose subsumption relation. Correspondingly, in order to be aligned with the hierarchically structured purposes, the tasks, which are designed to serve the purposes need to be hierarchically structured as well. Figure 5 shows an example of a task hierarchy with diagnosing as the super task, which is assigned the super purpose MT. Further, its two sub tasks General Check and Kidney Check are assigned sub purposes GT and KT respectively.

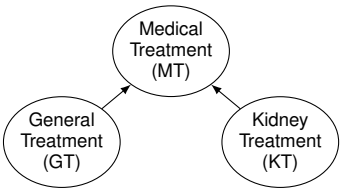

Figure 4. Example of a purpose hierarchy

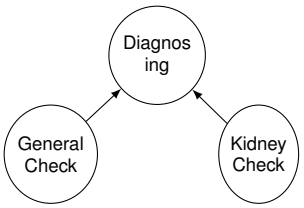

Figure 5. Example of a task hierarchy

Definition 9: (**Purpose of a task**) *For each task in an application, there exists a purpose in set P that is served by the task.*

$$\phi_T : T \rightarrow P$$

Where ϕ_T is a purpose assigning function for tasks and $\phi_T(t_i)$ is the purpose of the task t_i

Further, the authorization objects $\mathcal{O}_A$ of a system need to have specified purposes for which they are collected and stored. However, determining purposes for every RDF instances is time-consuming and error-prone. Hence, in this paper we consider that every OWL class in a domain of interest is assigned a set of usage purposes. The relationships R that are specific to the instances of a class and the instances themselves inherit the purposes of that class.

Definition 10: (**Purposes of OWL classes**) *Each OWL class in a domain of interest are assigned a non-empty set of purposes.*

$$\phi_C : C \rightarrow 2^P \setminus \emptyset$$

Where ϕ_C is a purpose assigning function for OWL classes and $\phi_C(c_i)$ are the purposes of the elements and the relationships R of the class c_i.

In the original Task-Based Privacy model a set of data transformation procedures are defined for each tasks. In order for a subject to perform its tasks it needs to be authorized to execute certain transformation procedure on the authorization objects. In the Semantic Web context however, query answering is the major essential operation [22]. Hence, we consider query answering as the only action by the subjects S on the authorization objects $\mathcal{O}_A$, in access control terms this represents the $READ$ action and denoted as $READ(\mathcal{O}_A)_S$.

B. Model Constraint and Properties

In this subsection, we formally define the security policy that constraints the behavior of a Semantic Web query answering system such that subjects only receive the information that they allowed to receive. Further, we define security

properties that need to fulfilled by the system in order to control the inference of specific information from a general set of information.

*Security Policy-1: (**S1**): A subject is granted read access to $\mathcal{O}_A$, only if the purpose of a subject's current task is contained in the set of purposes of the OWL classes that enclose the $\mathcal{O}_A$ or is a super purpose of a purpose in that set.*

$$\forall s_i \in S, \mathcal{O}_A \in G : READ(\mathcal{O}_A)_{s_i} \Rightarrow \phi_T(CT(s_i)) \in \phi_C(C(\mathcal{O}_A)) \vee \phi_T(CT(s_i)) \geq p_j, for\ p_j \in \phi_C(C(\mathcal{O}_A))$$

*Security Policy-2: (**S2**): The subject must be authorized to execute its current task.*

$$\forall s_i \in S : CT(s_i) \in AT(s_i)$$

The current task $CT(s_i)$ of s_i must be an element of its authorized tasks.

Security Properties: However, the soundness of the security policy (S1) is challenged by the interrelationships among the classes of O. In this sub section, we define security properties for OWL class taxonomies to ensure $S1$ is consistent. Properties for subclass, class union, and class intersection are defined.

Sublcass: Subclass axioms represent the taxonomy of OWL classes that describe the domain of interest K. Semantically, lower level nodes are more specific than the generic higher level nodes in a OWL class hierarchy [23]. In this context, to restrict access to specific subclasses, specific purposes need to be assigned to the subclasses. However, subclass axioms also imply membership of parent classes and this is an allowed inference. Hence, for the reasoning engine to include this inference, the purposes of the subclasses need to be included in the purpose assignment of their parent classes.

C1 (Subclass): Given the classes B, D and if $B \subseteq D$, then the purpose assigned to the subclass B must be included in the purposes assigned to its parent class D. If $B \subseteq D \Rightarrow \phi_C(B) \subseteq \phi_C(D)$.

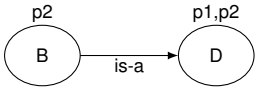

Figure 6. Example purpose assignments for subclass relationships

Figure 6 shows an illustrative example, where the subclass B is assigned the purpose p2 and that the purpose is included in the purposes for D. $S1$ implies that the task that is assigned the purpose p2 is allowed to view the instances of B and the semantic relation (implicit knowledge) that B is a subclass of D. Whereas, the task that is assigned the purpose p1 is allowed to view only instances of the more general class D.

Class Union: The union of two or more classes consist of instances that are member of at least one of those classes. Semantically, the union encompasses of instances of one or more specific subclasses. In this context, the purpose assignments of the specific subclasses that comprise an union class are included in the purpose assignments of that union class.

C2 (Class Union): Given the classes A, B, D and if A is a result of set $union$ of its subclasses B and D, then the purposes assigned to A must include the purposes assigned to its subclasses. If $A = B \cup D \Rightarrow \phi_C(A) \supseteq \phi_C(B) \cup \phi_C(D)$.

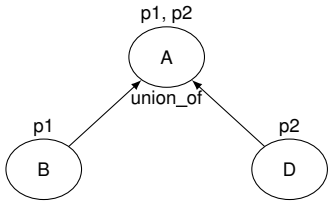

Figure 7. Example purpose assignments for class union relation

Figure 7 shows an illustrative example, where the union class A includes the purpose assignments p1, p2 of B and C respectively. According to $S1$ the tasks that are assigned either p1 or p2 are allowed to view the individuals of respective subclasses including the knowledge about the union.

Class Intersection: The intersection of two or more classes contains exactly every individual, which is a member of those concerned classes. Semantically, *i)* the overlapped class is more specific than the generic overlapping classes and *ii)* the overlapped class combines individuals belonging to two or more distinct classes which may have been collected for different purposes. Hence, access to the overlapped class must be constrained, so that unauthorized inference of the respective overlapping class is prevented as well as the illegal inference of overlapped class from the overlapping class. As consequence of the purpose hierarchy introduced in III-A, the purpose assignment of the overlapped class needs to dominate the purpose assignments of its overlapping classes.

C3 (Class Intersection): Given the classes B, D, E and if E is a result of set $intersection$ of B, D then the purpose assignment of E must dominate the purpose assignments of its overlapping classes. If $E = B \cap D$ then $\phi_C(E) \geq \phi_C(B) \wedge \phi_C(E) \geq \phi_C(D))$ and $\phi_C(E)$ is the lowest super purpose of B and D in the purpose hierarchy (i.e any other super purpose of B and D is dominating $\phi_C(E)$).

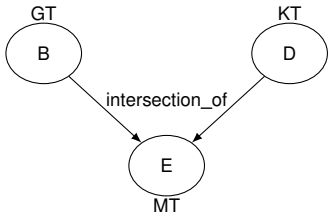

Figure 8. Example purpose assignments for class intersection relation

Figure 8 shows an illustrative example. The overlapping classes B and D are respectively assigned sub-purposes GT and KT from the example purpose hierarchy presented in Figure 4. In the Figure 4, it shows that MT is the super-purpose in the hierarchy that subsumes GT and KT. According to $S1$, the task

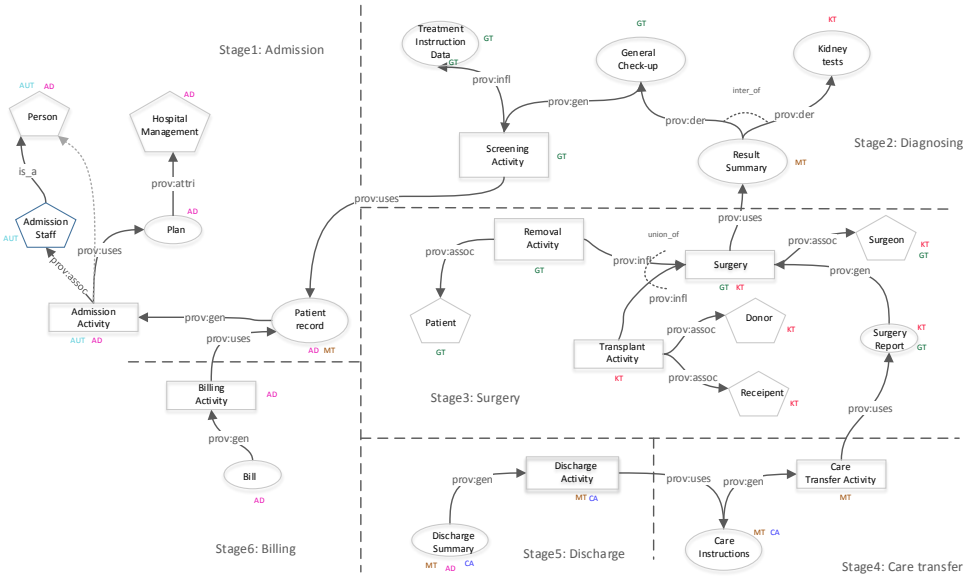

Figure 9. An example of a provenance OWL ontology for an imaginary surgical scenario

that is assigned MT which inherently subsumes the subtasks with purposes GT and KT is allowed to view the individuals of B, D and E. Whereas the subtasks with purpose assignments GT or KT is allowed to view the instances of the respective overlapping classes.

IV. Use Case: Access to Provenance Information in a Hospital Information System

An imaginary hospital called St.Mark hospital tracks provenance for the web documents to serve various provenance related automated queries. RDF is used to describe the provenance of web information using the ontology PROV-O. Access to such information needs to be managed for security, and especially for privacy reasons. The authorization policy of St.Mark hospital information is based on the purpose limitation principle. We consider the surgical part of the hospital service in this example. A patient needs to be admitted and diagnosed for a surgery procedure, hence the data usage purposes listed in Table I are identified for this scenario;

Table I. Purposes for data processing in a surgical care scenario

Purposes
Administration (AD)
Audit (AUT)
General Treatment (GT)
Kidney Treatment (KT)
Medical Treatment (MT)
Care Transfer (CA)

Graphical representation of the provenance ontology encompasses of OWL classes, taxonomy and relationships in a imaginary surgical scenario is shown in Figure 9

According to the PROV-Data Model (PROV-DM) [24], the ellipses represent the data items, the rectangles represent the processes or the activities that act on the data items and the hexagons represent the respective agents. Each of these PROV elements are modeled as OWL classes in Figure 9 including the respective relationships, which are also modeled in accordance with the PROV-DM. Each OWL class is assigned a set of purposes for which the data is collected and stored. In the following subsections we present a set of examples of, how the purpose limitation principle is enforced using the model described in Section III. In particular, we illustrated using an imaginary scenario, the security properties of the proposed model and the details of directly assigned and indirectly derived sub purposes.

A. Security Properties: Subclass and Union of Subclasses

In the OWL semantics, being member of a subclass implicitly means being member of a respective superclass. Hence, a query to an instance of a subclass includes the information that the instance is also a member of a superclass. Since super classes are more generic than the specific subclasses, the query to the superclass instances however, should not include the knowledge of the corresponding subclasses. Thereby, unauthorized inference of specific information from general information is prevented.

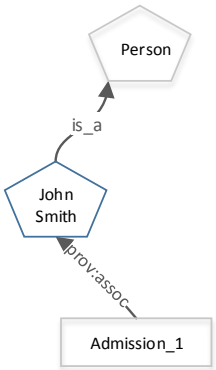

Figure 10. An example of a RDF subgraph that includes knowledge of respective super class

In Figure 9, "admission staff" class is more specific than the generic "person" class. According to $S1$, an auditor who is performing audit provenance query as part of an audit task to fulfill the audit purpose is authorized to read the RDF subgraph that is stored for audit purpose.

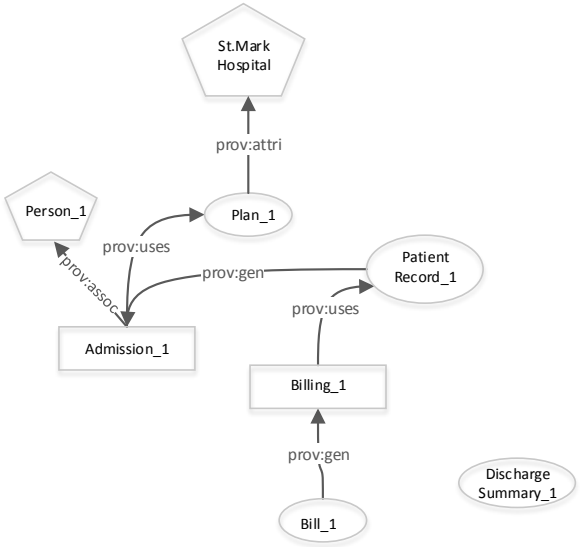

Figure 11. An example of a RDF subgraph pertaining to administration purpose

The resulting RDF subgraph is shown in Figure 10, which includes the implicit knowledge that "John Smith" who is an admission staff is also a member of a "person" class. Whereas, an administrator, who is executing a query that is related to the admission task for performing the administrative purpose, will receive the RDF subgraph depicted in Figure 11. This subgraph does not include the knowledge about a specific admission staff but the knowledge that the associated PROV agent, is a member of a "person" class.

The same principle is applied to the union of two or more subclasses. The classes that add up to a union are more specific than the generic union class. Querying for instances of subclasses that comprise the union includes the knowledge that the subclasses are part of an union class. Likewise, querying the instances of the union class includes the knowledge of the subclasses that comprise the union. This is because, unlike the generic super class that is an abstraction of infinite number of specific classes, the union class consist of exactly a finite number of subclasses.

In Figure 9, "Surgery" is a union of "Removal activity" and "Transplant activity" with purposes "GT" and "KT" respectively. Figure 12 shows an example of a returned RDF sub graph for a provenance query related to a general-check task that fulfills the general treatment purpose. The subgraph includes the knowledge that "Polyp removal" is part of an union class "Surgery".

B. Security Property: Class Intersection

In OWL class intersection, the overlapped class is more specific than the overlapping classes. Hence, it must be ensured that unauthorized inference of overlapped class from the generic overlapping class is prevented. Similarly, the inference from the overlapped class to the overlapping class need to be permitted because if a subject get hold of the all the overlapping classes then the subject can easily infer the overlapped class. This achieved in our model by means of the super-purposes and sub-purposes hierarchy. As result of the semantics behind the super-purpose, the overlapped classes need to be assigned super-purposes. Thereby, authorization on the overlapped class subsumes the purposes of the overlapping classes.

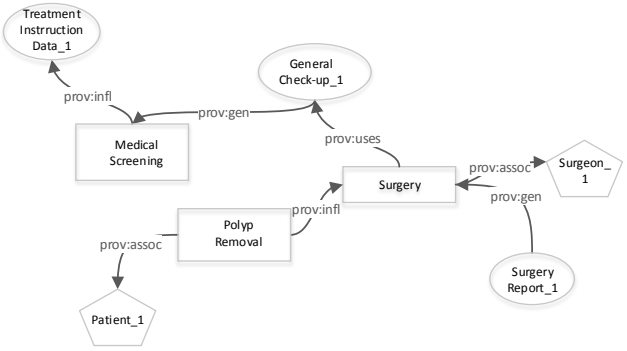

Figure 12. An example RDF subgraph for class union

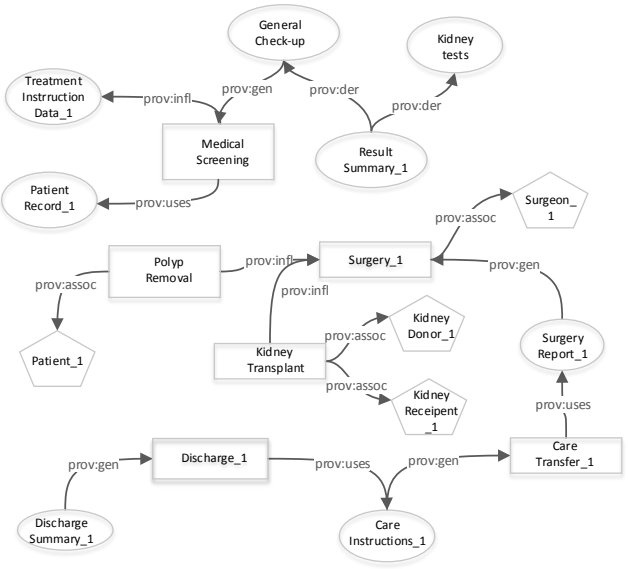

Figure 13. An example RDF subgraph that includes instances that are due to the indirectly derived purposes

In Figure 9, "Result Summary" is more specific than the generic overlapping classes "General Checkup" and "Kidney tests". A specialist executing a provenance query as part of the diagnosing task for the fulfillment of the medical treatment (MT) purpose, will receives the RDF subgraph shown in Figure 13. The subgraph does not only include the instance of the overlapping class "Result Summary" but also the knowledge that it is an intersection of classes "General Checkup" and "Kidney tests". This is due to the fact that the purposes of

both the "General Checkup" and "Kidney tests" classes are the sub purpose of MT. According to $S1$, a task, which is assigned a super purpose can access the information pertaining to its corresponding sub purposes. However, the RDF subgraph returned for the task that is assigned a sub-purpose GT does not include the knowledge of the overlapped class (see Figure 12).

V. Related Work

Much research on access control for provenance information does not acknowledge the emerging Semantic Web principles that underly the web provenance architecture. One exception is the work by Cadenhead et al. [25], they extend the access control language for generalized provenance model [26] with regular expressions. Regular expressions are used to identify relevant parts of the RDF graphs that represent provenance. Although, the policy specification of their model includes access purposes, they did not consider the entailments provided by OWL or RDF semantics.

Further, considerable amount of research has been devoted to investigate the access control models for RDF data stores. Jain et al. [13] propose a mandatory access control model for the RDF data stores that include derived RDF statements that follow from a RDF schema. However, OWL semantics that we have considered in our model are formally grounded and hence are more precise than the RDF schema. Furthermore, their model is based on a linear hierarchy of security classification labels assigned to the data objects, which is not pragmatic to define in the context of emerging applications except for the military domain. There are lot of research efforts on access control policy languages that specify access restrictions for semantically enriched information. Amongst which, the most relevant one is the work done by Kaushik et al. [12]. They propose a constraint logic based policy language to represent disclosure constraints for exposing parts of the ontology, and removing or desensitizing sensitive ontological concepts. However, in their model the disclosure constraints are not based on the access restriction attributes of the access objects, which is the primary focus of our model. On the similar basis, the work by Qin et al. [14] is an identity-based access control model rather than a mandatory access control model. Although, similar to our model their model is based on the relations between the OWL classes and how those relations can reveal information about one class from that of the other. We ascertain that both the work can be extended using our model.

Finally, significant amount of research effort has been put on automatic processing of privacy policies that enforce purpose limitation over the personal data access. Two main policy languages that formally represent enterprise data processing requirements are EPAL [27] and PPL [28]. Data usage restriction of these languages, however, are centered around data objects that does not represent semantic relationships. Similarly, Task-Based Privacy model [15] places technical controls for the implementation of legal privacy requirements such as purpose limitations but again the focuses is on the data objects that are not semantically enriched.

VI. Conclusions and Future Direction

The major strength of our model is that it recognizes the characteristics of the protection objects rather than the characteristics of the subjects. We ascertain however, that our model can be integrated into the role-based access control by authorizing the tasks to the roles instead of the subjects. In our model, we consider that the restriction attributes are for each OWL class including its relationships. However, the relationships that connect individuals of different classes might involve different type of semantics than class taxonomy hence may require a discrete access restriction attributes on its own. Furthermore, the abstraction on the OWL relationships due to our model introduces violation of integrity with respect to the OWL relationships. Figure 11 shows such a violation, where the instance "Discharge Summary_1" is unrelated to any class. Hence, in our future work we consider to study unauthorized inferences result from the OWL class relationships besides the OWL class taxonomy and consider to devise an abstraction mechanism for mitigating unauthorized inferences, which respect the OWL relationship constraints.

Acknowledgments

The authors are thankful to the SmartSociety, a project of the 7^{th} Framework Programme for Research of the European Community under grant agreement no. 600854, for funding the research that resulted in this publication. We extend our thanks to Rose-Mharie Åhlfeldt and anonymous reviewers for their invaluable feedbacks.

References

[1] 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," Commun. ACM, vol. 51, no. 4, Apr 2008, pp. 52–58. [Online]. Available: http://doi.acm.org/10.1145/1330311.1330323

[2] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-Science," Science, vol. 308, no. 5723, 2005, pp. 817–821.

[3] J. Cheney and R. Perera, "An Analytical Survey of Provenance Sanitization," CoRR, vol. abs/1405.5777, 2014. [Online]. Available: http://arxiv.org/abs/1405.5777

[4] S. B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen, "On provenance and privacy," in ICDT'11. ACM, 2011, pp. 3–10.

[5] U. Braun, S. Garfinkel, D. A. Holland, K.-K. Muniswamy-Reddy, and M. I. Seltzer, Provenance and Annotation of Data: International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, ch. Issues in Automatic Provenance Collection, pp. 171–183. [Online]. Available: http://dx.doi.org/10.1007/11890850_18

[6] U. Braun, A. Shinnar, and M. I. Seltzer, "Securing Provenance," in Proc. of the 3^{rd} Conf. on Hot Topics in Security, ser. HOTSEC'08, 2008, pp. 4:1–4:5.

[7] J. Reuben, L. A. Martucci, S. Fischer-Hübner, H. Packer, H. Hedbom, and L. Moreau, "Privacy Impact Assessment Template for Provenance," in Proc. of the Workshop on Challenges in Information Security and Privacy Management at the 11^{th} International Conference on Availability, Reliability and Security, August 2016.

[8] A. Syalim, Y. Hori, and K. Sakurai, "Grouping provenance information to improve efficiency of access control," in Advances in Information Security and Assurance, ser. LNCS. Springer Berlin Heidelberg, 2009, vol. 5576, pp. 51–59.

 ISBN: 978-1-61208-558-6

[9] A. Rosenthal, L. Seligman, A. Chapman, and B. T. Blaustein, "Scalable Access Controls for Lineage," in 1^{st} Workshop on the Theory and Practice of Provenance, ser. TAPP. USENIX, 2009.

[10] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and P. Yang, "Secure scientific workflow provenance querying with security views," in 9^{th} Int. Conf. on Web-Age Information Management, ser. WAIM, Jul 2008, pp. 349–356.

[11] T. Cadenhead, V. Khadilkar, M. Kantarcioglu, and B. Thuraisingham, "Transforming Provenance Using Redaction," in Proc. of the 16^{th} ACM Symp. on Access Control Models and Technologies, ser. SACMAT '11. ACM, 2011, pp. 93–102.

[12] S. Kaushik, D. Wijesekera, and P. Ammann, "Policy-based Dissemination of Partial Web-ontologies," in Proceedings of the 2005 Workshop on Secure Web Services, ser. SWS '05. ACM, 2005, pp. 43–52, ISBN: 1-59593-234-8, URL: http://doi.acm.org/10.1145/1103022.1103030 [accessed: 2018-04-08].

[13] A. Jain and C. Farkas, "Secure Resource Description Framework: An Access Control Model," in Proceedings of the Eleventh ACM Symposium on Access Control Models and Technologies, ser. SACMAT '06. ACM, 2006, pp. 121–129, ISBN: 1-59593-353-0, URL: http://doi.acm.org/10.1145/1133058.1133076 [accessed: 2017-04-08].

[14] L. Qin and V. Atluri, "Concept-level Access Control for the Semantic Web," in Proceedings of the 2003 ACM Workshop on XML Security, ser. XMLSEC '03, 2003, pp. 94–103, ISBN: 1-58113-777-X, URL: http://doi.acm.org/10.1145/968559.968575 [accessed: 2017-04-08].

[15] S. Fischer-Hübner, IT-security and Privacy: Design and Use of Privacy-enhancing Security Mechanisms. Berlin, Heidelberg: Springer-Verlag, 2001.

[16] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, vol. 284, no. 5, 2001, pp. 28–37.

[17] M. Lanthaler, D. Wood, and R. Cyganiak, "RDF 1.1 Concepts and Abstract Syntax," W3C, W3C Recommendation, Feb. 2014, URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ [accessed: 2017-04-08].

[18] T. Lebo, D. McGuinness, and S. Sahoo, "PROV-o: The PROV ontology," W3C, W3C Recommendation, Apr 2013, URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/ [accessed: 2017-04-08].

[19] B. C. Grau, P. Patel-Schneider, and B. Motik, "OWL 2 web ontology language direct semantics (second edition)," W3C, W3C Recommendation, Dec. 2012, URL: http://www.w3.org/TR/2012/REC-owl2-direct-semantics-20121211/ [accessed: 2017-04-08].

[20] B. Glimm and C. Ogbuji, "SPARQL 1.1 entailment regimes," W3C, W3C Recommendation, mar 2013, URL: http://www.w3.org/TR/2013/REC-sparql11-entailment-20130321/ [accessed: 2017-04-08].

[21] European Commission, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016.

[22] I. Kollia, B. Glimm, and I. Horrocks, "SPARQL Query Answering over OWL Ontologies". Springer Berlin Heidelberg, pp. 382–396, ISBN: 978-3-642-21034-1, URL: http://dx.doi.org/10.1007/978-3-642-21034-1_26 [accessed: 2017-04-08].

[23] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer (Second Edition)," Dec 2012, URL: https://www.w3.org/TR/owl2-primer/ [accessed: 2017-04-08].

[24] L. Moreau and P. Missier, "PROV-DM: The PROV Data Model," W3C, W3C Recommendation, Apr. 2013, URL: http://www.w3.org/TR/2013/REC-prov-dm-20130430/ [accessed: 2017-04-08].

[25] T. Cadenhead, V. Khadilkar, M. Kantarcioglu, and B. Thuraisingham, "A Language for Provenance Access Control," in Proc. of the 1^{st} ACM Conf. on Data and Application Security and Privacy, ser. CODASPY '11. ACM, 2011, pp. 133–144.

[26] Q. Ni, S. Xu, E. Bertino, R. Sandhu, and W. Han, An Access Control Language for a General Provenance Model. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 68–88, ISBN: 978-3-642-04219-5, URL: http://dx.doi.org/10.1007/978-3-642-04219-5_5 [accessed: 2017-04-08].

[27] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise privacy authorization language (EPAL 1.2)," Submission to W3C, vol. 156, 2003.

[28] PrimeLife, "PrimeLife – Privacy and Identity Management in Europe for Life: Policy Languages," Available at http://primelife.ercim.eu/images/stories/primer/policylanguage-plb.pdf, 2011.

 ISBN: 978-1-61208-558-6