Semantic-based Privacy Protection of Electronic Health Records for Collaborative Research

Yang Lu & Richard O. Sinnott
Dept. of Computing and Information System
University of Melbourne
Melbourne, Australia
luy4@student.unimelb.edu.au

***Abstract*—Combined health information and web-based technologies can be used to support healthcare and research activities associated with electronic health records (EHRs). EHRs used for research purposes demand privacy, confidentiality and all information governance concerns are addressed. However, existing solutions are unable to meet the evolving research needs especially when supporting data access and linkage across organization boundaries. In this work, we show how semantic methods can aid in the specification and enforcement of policies for privacy protection. This is illustrated through a case study associated with the Australasian Diabetes Data Network (ADDN), the national paediatric type-1 diabetes data registry and the Australian Urban Research Infrastructure Network (AURIN) platform that supports Australia-wide access to urban and built environment data sets. Specifically we show that through extending the eXtensible Access Control Markup Language (XACML) with semantic capabilities, we are able to support fine-grained privacy-preserving policies leveraging semantic reasoning that is not directly available in XACML or other existing security policy specification languages.**

Keywords-privacy; electronic health records (EHRs); ontology; semantic web rule language (SWRL); XACML; obligation component

I. INTRODUCTION

Online data management is essential in the modern digital-rich world. In the biomedical domain, there is a seismic change occurring from paper-based documents to electronic records. Stakeholders across health industries have appreciated significant benefits through increased digitization including information accuracy and efficiency. Many challenges remain to be addressed however. Clinical and biomedical data are obviously sensitive since they often contain individual personal information. To protect personal and sensitive data from malicious or non-malicious leaks, organizations have adopted a range of solutions, such as requiring informed consent from data subjects, removing identifiable information and data anonymisation. Technical security approaches such as authentication, authorization, auditing and accounting can help support such activities. Authentication is used to identify legitimate users often by checking their usernames and passwords; authorization further restricts the operations that can be performed by an authenticated user, whilst auditing and accounting capabilities are used to record a history of access and usage. This work predominantly focuses on authorisation – the most demanding of capabilities that are required. Different access control models and policy languages have been defined to support authorisation. The international standard, XACML (eXtensible Access Control Markup Language) includes common request/response protocols and associated policy components to support authorisation [1]. Due to its standardisation, widespread adoption and rich expressiveness, XACML has been widely implemented in EHR systems [2] [3] [4]. In particular, the obligation component of XACML has the potential to seamlessly introduce other policies (e.g. privacy policies) by specifying what extra actions should be taken before data is released or linked. However traditional XACML policies are specified in a static manner, which cannot fit the demands of distributed application due to the heterogeneity. For that reason, we propose to extend access control policies with semantic functionalities, building on previous work [5]. Through reasoning on the domain knowledge with semantic capabilities, we show how we can overcome syntactic language barriers and support advanced authorisation of subjects involved in a given collaboration. One challenge in undertaking this is where multiple datasets are combined and exchanged with heterogeneous policies. Due to the information accumulation, it is often not possible to provide adequate protection by enforcing static and independent policies on data linkages across organisation boundaries [6]. To solve this, we present a semantic approach to mitigate and reason about potential privacy leakage. Through reasoning on the domain knowledge, we can reduce the risk of privacy leakage that can occur in inter-organisational linkage scenarios.

The rest of the paper is structured as follows. Section 2 reviews the state of art in related fields including data anonymisation, access control models and semantic technologies. Taking the type-1 diabetes patient data from ADDN as an example, Section 3 explores how to formulate XACML policies into semantic concepts so as to fit the demands of current applications. Particularly, the privacy requirements are specified as obligations related to privacy concerns. In Section 4, a linkage case study is introduced that utilises geospatially-annotated data from ADDN and the AURIN platform. Instead of adding new rule/policies for linkage, we show that through a semantic approach we can

2324-9013/16 $31.00 © 2016 IEEE
DOI 10.1109/TrustCom.2016.104

deliver real-time protection and reuse of existing policies. Section 5 summarises conclusions and identifies areas of future work.

II. Related Works

Data sharing within clinical collaborations is essential. In the field of healthcare, sharing data for secondary use allows discoveries in clinical trials for new drugs, treatments, benchmarking care and outcome more generally. Since health data contains personal information, researchers are expected to use data in both an ethical and confidential manner. There are several essential procedures that are required to support this research process [7]:

- *Consent*: data related to an individual's health typically contains both identifiable and non-identifiable information [8]. Before granting the access to this data for research purposes, data holders (hospitals/clinical institutions) are often required to obtain consent from patients (or assent in the case of minors or others that are not legally able to provide consent). Upon receiving confirmation of consent, data holders are then better equipped to allow secure queries for research use outside of the given hospital setting – noting that consent alone is not enough to allow access.
- *Anonymity*: to protect patient privacy, health data for secondary use often needs to be de-identified (anonymized). Depending on the purpose and nature of the research, different degrees of data anonymisation are often demanded. For instance rare diseases may have specific restrictions on identify data, e.g. year of birth only, compared to more common conditions.
- *Access control*: data needs to be protected by access control policies to ensure ethical and validated data access, as well as support minimal knowledge leakage. Domain-specific regulations and risk management need to be defined and enforced to ensure this process is adhered to. These rules are often formalized based on minimizing potential security issues [3] [9].

A. Data Anonymisation

De-identified EHRs are often shared for research purposes. In Australia, the National Statement on Ethical Conduct in Human Research has classified medical data into individually identifiable, re-identifiable and non-identifiable. In addition, it specifies in what contexts, which type of data is allowed to be collected, stored and published. In the United States, privacy issues are covered by Health Insurance Portability and Accountability Act 1996 (HIPPA) which defines different levels of de-identification as guidelines of data anonymity. A common model to achieve this is through "safe harbors" whereby de-identification can be achieved by removing 18 commonly identifying attributes such as name, address, date, biometric information, serial numbers of personal devices etc. With the increasing complexity in data usage, however, it is not always easy to group data items according to their sensitivity. What is more, isolated datasets can often reveal sensitive information. For instance, De Montjoye has shown that only four pieces of spatio-temporal information are enough to re-identify individuals [10]. Moreover, by matching clinical data with external resources, hidden factors/trends about patients/diseases can be ascertained.

Technical solutions based on k-anonymity aim to reduce the risk of re-identification by obscuring individuals with other k-1 identical records [11] [12]. To achieve this, the approach relies on generalising or suppressing "quasi-identifiable" variables. To further strengthen the power, k-anonymity variants such as t-closeness [13] and l-diversity [14] are proposed to assimilate value distributions and reduce the granularity of data representation. In addition, the k-anonymity algorithm can be further extended with specific application purposes. For instance, in addition to a global measure k, [15] deems that the risk evaluation should include the trustworthiness of users. In [16], data owners are allowed to express their privacy preferences and then influence the overall anonymisation. Another important practice is differential privacy, which aims to maximize the accuracy of queries from statistical databases while minimizing the chance of re-identification [17]. However, it may not always be the case that it is possible to identify differential privacy distributions in inter-organisational data linkage and data sharing applications, especially when dealing with arbitrary queries on diverse data models. Through linking data, it is possible to infer new knowledge about individual (or group) that cannot always be predicted. As a consequence, data cannot be protected by independent solutions. As a result, such techniques do not fulfil the unique needs for healthcare in data linkages across multiple and potential dynamic/evolving collaborations.

B. Access Control and Semantic Technoogy

Restricting access to datasets to meet associated regulations is essential. Role-Based Access Control (RBAC) is a typical access control model. It was originally proposed to provide secure access based on grouping users/resources and attaching security levels to these groups (roles/clearances). In addition to the role concept, attribute based access control (ABAC) incorporates more extensive attributes into policies, and thus enables a finer-grained access control. XACML has been widely used to implement ABAC systems. Variants of ABAC models designed for special applications also exist, e.g. factoring in location and time as additional conditions that can be used to make decisions about the authorisation [18-20]. The XACML framework supports an obligation component that can be used to define the actions that must be taken when permissions are granted [21]. In other words, the permissible actions cannot take effect until the associated obligations are executed successfully. For instance, the policy "a clinician can collect and disclose personal information from a child by obtaining parental consent from the child's guardians" can be constructed by a permission (collect and release a child's details) and an obligation (obtaining consent from the child's guardians). This can be specified as a premise and a consequence respectively. As part of an access control

framework, obligations can facilitate the implementation of privacy regulations even if they are not explicitly devised for that purpose. For instance, [22] demonstrates an enforcement of HIPPA privacy rules within access control models in healthcare applications. In [23], further methodologies are provided in devising privacy-aware systems through turning data protection regulations as obligations and associated actions.

As mentioned, privacy leakage happens in isolated/combined datasets while semantic reasoning is one way to deliver dynamic protection. Language barriers in decentralised systems can often be tackled by formalising subsumption and equivalence through the Ontology Web Language (OWL) knowledge base [24-26]. In dealing with complex situations with various regulations, it is often required to have policies generated based on implicit knowledge. The Semantic Rule Web Language (SWRL) [27] can be adopted to support policies based on domain knowledge [28-30]. Researchers in this field have explored semantic methods by giving most attentions to facilitating authorisation; however privacy leakage remains a key challenge for distributed applications. To fill this gap, we semantically extend XACML with extra obligation measures [31]. Through using the ontology editor, protégé 4.0 (http://protege.stanford.edu/), policy components are conceptualised as classes, instances, properties as well as semantic rules. Finally, we take the linkage scenarios of two major and real-life systems, ADDN and AURIN to illustrate the way interconnection between authorisation and privacy protection on ad hoc data linkage queries can be realised.

III. SEMANTIC PRIVACY AWARE XACML

A. *Authorisation Framework*

The XACML policy framework consists of three levels of entities: the **PolicySet**, the **Policy** and the **Rule**. Specifically, a **Rule** represents the smallest unit specifying access requirements. Each rule is assigned with an **Effect**, e.g. deny, permit or not applicable, which becomes effective when the (requests) attributes are validated against the (policies) constraints of the **Target** and **Condition**. In XACML, a **Target** builds the fundamental constraints for the context of **Subject**, **Action**, **Resource** and **Environment**. Optionally, there are some higher-level constraints enclosed in the **Condition** to further restrict authorisation. Beyond compliance checking, policy makers have rights to add extra security measures in **Obligation** as the "last defender", e.g. special privacy regulations on data use. TABLE I shows the semantic specification of the XACML framework.

TABLE I. MAPPING XACML COMPONENTS INTO OWL CONCEPTS

Class	Associated Properties	Range
Policy	hasRule hasObligation	Rule Obligation
Rule	hasTarget hasEffect fromPolicy	Target Effect Policy
Target	hasSubject hasResource hasAction hasEnvironment	Element
Element	hasConstraint	Attribute

In addition to the policy components and their explicit relations, compliance checking is typically based on concrete facts. For instance, a pseudo policy (Policy-a) in Figure 1 defines *"clinicians can collect and read the type-1 diabetes mellitus (T1DM) patients records for research by executing the de-identification operation"*. The constraints "*Clinician*", "*T1DMPatient*", "*Read*" and "*ForResearch*" are respectively applied on the elements **Sub_a**, **Res_a**, **Act_a** and **Env_a**, which collectively construct the target and rule, i.e. **Tar_a** and **Rule_a**. Besides, the obligation "De-identification" is the premise of the **permit** decision becoming effective in the system. For instance, Figure 2 shows the OWL modelling the instance of policy, rule and target.

```
<Policy Id="Policy_a" Algorithm="deny-unless-
permit">
  <Rule Id="Rule_a" Effect=Permit>
   <Target Id = "Tar_a">
   <Subject Id="Sub_a" AttributeValue
="Clinician"/>
   <Resource Id="Res_a" AttributeValue
="T1DMPatients"/>
   <Action Id="Act-a" AttributeValue ="Read"/>
   <Environment Id="Env_a" AttributeValue
="ForResearch"/>
   </Target>
   </Condition>
  </Rule>
  <Obligation Id = "De-identification"
FulfillOnEffect = "Permit"/>
</Policy>
```

Figure 1. Example XACML policy (Policy-a)

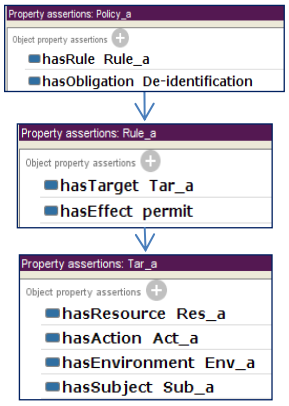


Figure 2. OWL-based policy specification (Policy-a)

In addition to using the OWL, semantic rules can also be used to realise XACML compliance checking, which helps reduce the efforts on specification by reusing domain-specific contents. As shown in Figure 2, the attribute "T1DMPatients" is described under the resource tag, and thus it only acts as a resource attribute in Tar_a, Rule_a and Policy_a. If it needs to be applied in other contexts, e.g.

Tar_b, Rule_b and Policy_a (the nesting structure allows one policy containing more than one rule), it typically has to be stated again. However, through reasoning on the Semantic Rule (1) and the semantic model, the "context" information can be dynamically assigned without repeated statements. As shown in Figure 3, based on the explicit relations of target elements and relevant rules, Tar_a can be dynamically associated with constraints in different aspects.

Target(?a), hasResource(?a, ?b), hasConstraint(?b, ?c) → hasResourceConstraint(?a, ?c) (1)

In addition to target elements, rules can be formulated based on their (semantic) dependency, and use in policies and policy sets. For example, through reasoning on the Semantic Rule (2), the subject constraint "*Clinician*" can be associated to the related rules (Rule_a).

Rule (?a), hasTarget (?a, ?b), hasSubjectConstraint(?b, ?c) → hasSubjectConstraint (?a, ?c) (2)

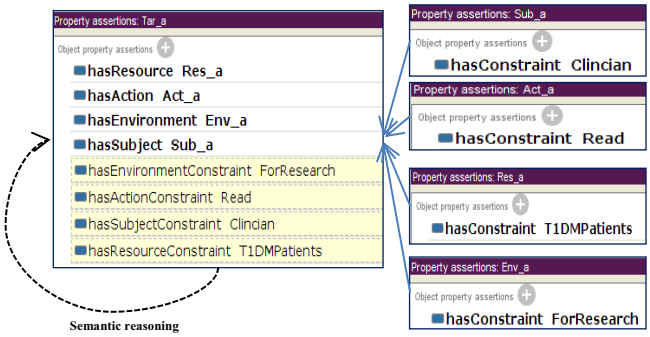


Figure 3. Semantic reasoning on Tar_a constraints

Similar to policies, the XACML request protocol is described with attributes of different aspects [32]. Correspondingly, they are mapped to semantic properties "*has_X_category*" and distinct attributes. Suppose the *Request_a* is submitted to the system like *"clinicians (subject) who are part of Project-T1DM (environment) request to read (action) the T1DM patient records (resource)*". To show the feasibility in compliance checking, this can be modelled as certain semantic concepts in the TABLE II. Different from policy management, it is unnecessary to formulate requests by semantic reasoning since technically they can be arbitrary statements that originate from external parties.

TABLE II. MAPPING XACML REQUEST PROFILE INTO OWL CONCEPTS

Class	Instances	Associated properties	Instances
Request	Request_a	has_Sub_Category	Clinician
		has_Obj_Category	T1DMRecords
		has_Act_Category	Read
		has_Env_Category	Project-T1DM

Policy evaluation is separated into a two-stage reasoning process. At the first stage, candidate rules pairs are identified through evaluating on the "match" function. It is noted that only the attributes falling into the same contexts are compared to make authorisation decisions. For instance, Semantic Rule (3) shows that through applying the "pairwise properties" *hasEnv_Category - hasEnvironmentConstraint* the consistency of attributes can be ensured. In the same way, rules on attribute comparison can also be formulated for the contexts of Subject, Resource and Action.

Rule (?a), Request (?b), hasEnv_Category (?b, ?su), hasEnvironmentConstraint (?a, ?sub), match (?su, ?sub) → candidateRuleE (?b, ?a) (3)

Attributes (such as literals, numeric, time, date) can be extensively covered through extending the definition of *match()*. For instance, with the built-in property *sameAs* identifying the same entities (instances/classes), Rule (4) can enable the *match()* function to compare identical semantic concepts [33]. A more complicated case happens among the numeric values, such as ">10". To support this, another built-in function, *greaterThan*, as well as the data property *isValuedAs (?x, ?value)* can be used to take numeric attributes, as shown in Rule (5). To a broad extent, *match()* can be formulated for particular applications through attaching domain knowledge. For instance, the environment attribute *Project_T1DM* (*Request_a)* refers to a project where the clinician is involved. For example, if there is background knowledge in the hospital that people have the same purposes as the projects they are working on, this can be expressed in Rule (6), i.e. research project *Project_T1DM* and *hasPurposeOf(Project_T1DM, ForResearch)*, through reasoning about the knowledge and rules (3) and (6), it is possible to deduce the intermediate result *match(Project_T1DM, for_research)* and then recognise the candidate rule, i.e. *candidateRuleE (Request_a, Rule_a)*.

Attribute(?x), Attribute(?y), sameAs(?x, ?y) → match(?x, ?y) (4)

Attribute (?a), Attribute(?b), isValuedAs(?a, num1), isValuedAs(?b, num2), greaterThan (num1, num2) → match (?a, ?b) (5)

Attribute (?a), Attribute (?b), hasPurposeOf (?a, ?b) → match (?a, ?b) (6)

Policy compliance requires that all target attributes that are included in the rule are satisfied by the request. Therefore, based on the searched candidate rules, it is necessary to determine the applicable rules by combining intermediate results of the last stage. For that reason, Semantic Rule (7) implies that the compliance between XACML rules and requests should be based on attributes covering all aspects.

Rule (?a), Request (?b), candidateRuleS (?b, ?a), candidateRuleR (?b, ?a), candidateRuleE (?b, ?a), candidateRuleA (?b, ?a)→ applicableRule (?b, ?a) (7)

Each XACML rule should have one effect assigned. Due to the nesting structure, it is possible for a single request to have multiple rules that need to be applied. Therefore, to achieve a final decision, XACML standardises a set of rule combining algorithms [34]. Since the open-world assumption (OWA) of the semantic web assumes incomplete information is not equal to false or failure [35], "permit" is defined as the default effect in this semantic model. For this reason, the algorithm "deny-unless-permit" is adopted to enforce the effects as shown in Rule (8). In particular, *hasRule* and *fromPolicy* are inverse properties, which means the expression *fromPolicy(?a, ?b)* is semantically equivalent to *hasRule(?b, ?a)*. With this formulation, Rule (8) can be used to identify applicable rules associating with the same policy.

Request(?x), applicableRule(?x, ?a), hasEffect(?a, permit), fromPolicy(?a, ?c) → finalEffect(?x, permit), finaEffectFrom (?x, ?c) (8)

In the healthcare field, having extra security measures is essential because it allows privacy policies to be implemented on certain data points while avoiding the answers based upon "zero" or "all" permissions. In XACML policies, the obligation component is the key to achieving this goal. Taking the result such as *finalEffect(Request_a, permit)* and *finalEffectFrom(Request_a, Policy_a)* it is possible to locate the attached obligations. Rule (9) shows how to associate obligations with the request. According to the definition of *Policy_a*, the inferred obligation *De-identification* should be implemented before the permission to *Request_a* takes effect.

Request(?x), finaEffectFrom(?x, ?y), hasObligation(?y, ?z) → hasObligation (?x, ?z) (9)

So far we show that semantic methods can facilitate policy development and evaluation. As shown in Figure 4, attributes are the elements that underpin the matching between requests and policies. With a set of candidate attributes and reasoning (Stage 1) the syntax of the information can be checked; applicable rules are then located in Stage-2, resulting in effects and obligations that are made to ultimately make access decisions.

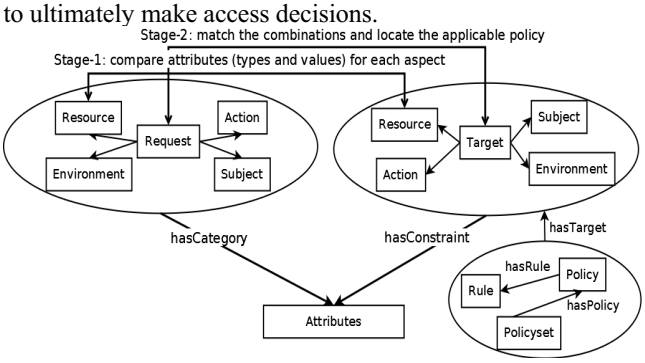


Figure 4. Evaluate policy and request in parallel

B. Privacy Protection – A case study based on ADDN

In clinical systems, the contents of single and combined databases demands various levels of sensitivity and privacy are adhered to. Through reasoning on Semantic Rules (7)-(9), authorising decisions can be made with certain obligations enforced. We consider here a scenario associated with the Australasian Diabetes Data Network (ADDN - http://www.addn.org.au/). ADDN provides a centralized repository of patients with type-1 diabetes across Australia. ADDN supports an RBAC model (Clinicians, Coordinators and Researchers) with specific levels of access (Centre, Multiple Centre and All) that are used to restrict the access to and use of medical data by researchers. As shown in Figure 5, ADDN records can be expressed in a patient-centered pattern. Since policies are formulated using the data schema, patient records can be represented at the structural level, i.e. *Patient(Patient_x), hasType(Patient_x, Types), hasEthnicity(Patient_x, Ethnicities)* and *hasZIP(Patient_x, ZIPs)*. The actual database in ADDN has collected over 200 data points about patients with over 5000 patients currently entered. To address privacy protection concerns, we consider the access to and use of the data related to the disease type, ethnicity and zip code.

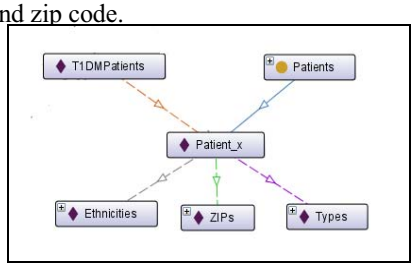


Figure 5. Semantic ADDN data schema (Patient_x)

In ADDN, using the aforementioned data elements, the risk of privacy leakage is reduced by generalising/supressing identifiable attributes. For instance, ethnicity information is specified by referencing the Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG)[1], as shown in TABLE III. Specifically, it is organised in a 3-level structure, such as "6 North-east Asian", "61- Chinese Asian" and "6101-Chinese". Considering the health conditions identified from certain groups or places may cause stigmatization and discrimination [36], hence one privacy demand is that only the aggregated ethnicities (at the first level of obfuscation) should be released from ADDN.

TABLE III. HIERARCHICAL STRUCTURE OF ASCCEG

Broad group	Narrow group	Cultural and ethnic group
6 North-East Asian	61 Chinese Asian	6101 Chinese 6102 Taiwanese 6199 Chinese Asian, nec.
	69 Other North-east Asian	6901 Japanese 6902 Korean 6903 Mongolian

It might seem that no privacy leakage could happen however this may not be the case in particular places. As shown in Figure 6, the Census statistics from the Australian Bureau Statistics (ABS) indicates the ethnicity distribution in the Box Hill (a suburb of Victoria, Australia) is comprised of more than 20 percent Chinese[2]. As a result, there is an increasing chance to identify which records are from Chinese patients because the combination of "North-East Asian" and "3128" (the zip code of Box Hill)[3] is almost equivalent to "Chinese residents in Box Hill". Since ADDN should avoid potential re-identification based on ethnical details, a specific obligation (e.g. de-identification) can be defined to block certain combinations of queries and results.

[1] Australian Bureau Statistics (ABS). Australian Standard Classification of Cultural and Ethnic Group (2011). http://www.abs.gov.au/ausstats/abs@.nsf/mf/1249.0

[2] Box Hill, Victoria. Demographics. https://en.wikipedia.org/wiki/Box_Hill,_Victoria#Demographics

[3] Victoria Postal Codes. http://www.geonames.org/postal-codes/AU/VIC/victoria.html

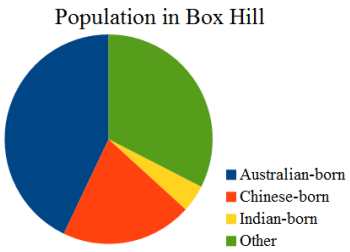


Figure 6. Partial statistics of Box Hill residents

Queries need to be authorised. Through semantic reasoning, only the data in the request should be used to answer the query. Therefore, Rule (10) can be formulated to verify arbitrary queries through the expressions with *hasObligation* and *queryFrom*. The service *Mask* refers to the concrete implementation of *De-identification*. According to the privacy policies, this can detect the risk of privacy leakage and then properly process the data with "sensitive" attributes. Only the permitted resource will be used to answer the query. For instance, the resource in question "*PatientT1DM*" refers to the EHRs collected from T1DM patients. As mentioned, *Patient_x* in the data schema can represent all records. This can be refined by the specific attributes. For instance, taking an aggregated zip code, 3XXX, we can focus on a subset of patients. Through reasoning on the previous results, the consequence can indicate the target records (answerQuery(*Patient_x, Query_a_n)*) and services (*Mask(Patient_x)*).

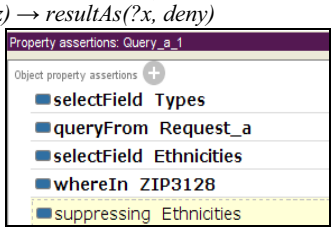
Query (?q), Request (?r), hasObligation(?r, De-identification), hasRes_Category(?y, ?d), hasPatient (?d, ?p), queryFrom(?q, ?r) → Mask(?p), answerQuery(?p, ?q) (10)

Based on those intermediate results, Semantic rule (11) - (13) are used to conduct the risk analysis. Particularly, the class **RiskScope** is formulated to indicate sensitive attributes such as "3128" and "North East Asian" in this case. In answering the query with such attributes as conditions and results, e.g. "`SELECT …, Ethnicity; FROM ADDN; WHERE ZIP= '3128'`", the resulting dataset should be suppressed/generalised since the condition can make the privacy information (i.e. specific ethnicity) inferable, depending on external knowledge. The first query case is shown in Figure 7 a). The ethnicity values are suppressed due to aggregate contents found in the system. Similarly, to the query such as "`SELECT …, ZIPs; FROM ADDN; WHERE Ethnicity='North East Asian'`", zip codes in the result set can be generalised like "312X", as shown as Figure 7 b). Considering the high risk of releasing ethnicities and zip codes, queries such as "`SELECT …, ZIPs, Ethnicities…`" should be directly denied, as shown in Figure 7 c).

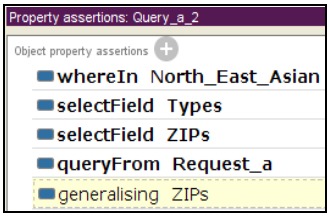
Mask(?x), Ethnicity(?e), ZIP(?z), RiskScope(?z), answerQuery(?x, ?q), Query(?q), selectFields(?q, ?e), whereIn(?q, ?z) → suppressing(?q, ?e) (11)

Mask(?x), Ethnicity(?e), ZIP(?z), answerQuery(?x, ?q), Query(?x), RiskScope(?e), selectFields(?q, ?z), whereIn(?q, ?e) → generalising(?q, ?z) (12)

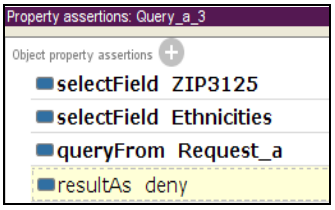
Query(?x), ZIP(?z), Ethnicity(?e), selectFields(?x, ?e), selectFields(?x, ?z) → resultAs(?x, deny) (13)



a) Query_a_1 for patients' ethnicities



b) Query_a_2 for patients' zip codes



c) Query_a_3 for patients' zip code and ethnicities

Figure 7. Example queries from request_a

Using domain-specific knowledge, semantic rules can be reasoned to compute these extra measures related to releasing query results. For instance, both *generalising (Query_a_2, ZIPs)* and *suppressing (Query_a_1, Ethnicities)* can be achieved through a semantically-enabled policy decision point (PDP) and then enforced by a policy enforcement point (PEP) related to the queries. Since the purpose here is to protect linkage privacy, such local policy constructs provide the foundation for policy generation.

IV. Case Study- Privacy Protection on Linkage AURIN and ADDN

In addition to querying a single dataset, linkage queries may also lead to privacy threats. To show the semantic approach is feasible for linkage applications, we consider a scenario where a researcher wishes to find the correlation between alcohol consumption and T1DM incidence. In doing so, it is necessary to compare the number of T1DM patients and their neighbouring bottle shops. The Australian Urban Research Infrastructure Network (AURIN, http://aurin.org.au/) was established as a comprehensive research platform to support seamless and secure access to a wide array of data. It includes over 2000 data sets from 70 major and typically definitive data agencies. This includes the official details of licensed premises provided by the Victorian Commission for Gambling and Liquor Regulation

(VCGLR)[4]. As shown in Figure 8, a typical scenario can be depicted where the Zip codes of both data sets facilitates the data linkage of T1DM and locations of bottleshops. In this case, we consider Alice (Clinician) queries the ADDN data with "bottle-shop data". TABLE IV shows the linked dataset of ADDN and AURIN. Since the accumulating contents may distinguish some aggregated values, it is necessary for the linking component to extend the security measures before releasing any results.

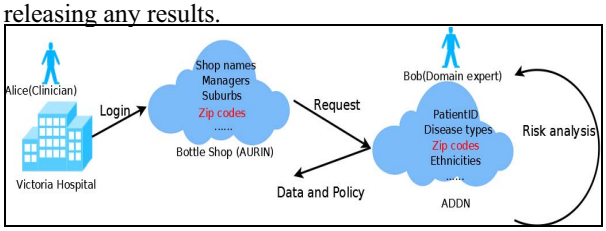


Figure 8. Access to linked datasets from ADDN and AURIN

TABLE V shows the request to the query such as "SELECT Suburbs, ZIPs; FROM Linkage; WHERE Ethnicity ='North East Asia'". According to Rule (12) in ADDN, sensitive zip codes will be aggregated like "312X", "31XX", "3XXX" etc. However, newly-added variables of Suburb may make the protection ineffective since the suburb-postcode mappings are publicly accessible. Based on the existing classification system[5], both Australian zip codes and suburbs can be semantically modelled and associated to each other through the expressions such as *subclassOf("3128","312X"), subclassOf("3129","312X")* and *sameAs("3128", "Box Hill")* etc. By reasoning about the extended knowledge and existing semantic rules, the resulting data can be organised as TABLE VI with decreased risk. In addition, attribute combinations such as "language-nationality", "gender-year of birth" may also pose threats to linkage applications. Through using the semantic model, privacy policies can be dynamically generated by reasoning about the linked knowledge from independent sites. Compared with adding particular policies for each query, this work can satisfy the demands of distributed systems by delivering dynamic protection to ever-changing resources.

TABLE IV. LINKAGE DATABASE BASED ON ZIP CODES

SID*	Type	Ethnicity	Zip code	Shop ID*	Suburb
099999	1	North East Asia	3128	31204487	Box Hill
099999	1	North East Asia	3128	31215064	Box Hill
099999	1	North East Asia	3128	31249712	Box Hill
099998	1	South East Asia	3053	32233992	Carlton
099998	1	South East Asia	3053	31200296	Carlton
099998	1	South East Asia	3053	31921796	Carlton
099998	1	South East Asia	3053	32227111	Carlton
……	……	……	……	……	……

TABLE V. QUERYING RESULTS BY ETHNICITIES*

SID*	Zip code	Suburb	Ethnicity
099999	312X	Box Hill	North East Asia
099999	312X	Box Hill	North East Asia
099999	312X	Box Hill	North East Asia
……	……	……	……

*query conditions in dash lines are not included in the result

TABLE VI. QUERYING RESULTS BY ETHNICITIES (AFTER REASONING)

SID*	Zip code	Suburb	Ethnicity
099999	312X	Box Hill, Camberwell, Mont_Albert_ North, Richmond, Burnley, Cremorne……	North East Asia

Techniques such as k-anonymity and XACML are insufficient to tackle arbitrary queries in linkage applications due to their static nature. Specifically, the k-anonymity model requires all the query results meet a uniform criteria such as "the equivalence class size should be at least equal to 3" in the case where data is only released if there are 3 or more individuals in a given suburb for example. However, in answering the query shown in the last section, it is possible to violate the regulation by releasing the contents as shown in TABLE V. Compared with traditional access control policies, this work can not only connect the authorization and privacy protection through semantic inferences, but also facilitate the dynamic demands of knowledge expansion in linkage systems. It is noted that this work is not to replace existing techniques. Instead, it is intended to complement the deficiencies of distributed applications.

V. CONCLUSIONS

In this work, we present a semantic XACML model to authorize access to sensitive information during data linkage. Through extending XACML policies with semantic rules, we have shown that we are able to protect private information of patients across organizational boundaries. In the next stage, we are going to further explore trust management and integrate it into the access control system to support richer sensitive information scenarios in the health domain.

ACKNOWLEDGMENT

The ADDN project is funded by the Juvenile Diabetes Research Foundation. The AURIN project is funded by the Department of Education of Australia. We gratefully acknowledge their support.

[4] Victorian Commission for Gambling and Liquor Regulation (VCGLR). www.vcglr.vic.gov.au

[5] Australian Statistical Geography Standard (ASGS). http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistica+Geography+Standard+(ASGS)

REFERENCES

[1] A Brief Introduction to XACML. https://www.oasis-open.org/committees/download.php/2713/Brief_Introduction_to_XACML.html

[2] Milutinovic, S. (2008). The need for the use of XACML access control policy in a distributed EHR and some performance considerations. Medical and Care Compunetics 5, 137, 346.

[3] Zhang, R., Liu, L., & Xue, R. (2014). Role-based and time-bound access and management of EHR data. Security and Communication Networks, 7(6), 994-1015.

[4] Bhartiya, S., Mehrotra, D., & Girdhar, A. (2015). Proposing hierarchy-similarity based access control framework: A multilevel Electronic Health Record data sharing approach for interoperable environment. Journal of King Saud University-Computer and Information Sciences.

[5] Lu, Y., & Sinnott, R. O. (2015). Semantic Security for e-Health: A Case Study in Enhanced Access Control. 12nd IEEE International Conference on Advanced and Trusted Computing (ATC 2015), Beijing, China, in press.

[6] Willey, H., Eastwood, B., Gee, I. L., & Marsden, J. (2016). Is treatment for alcohol use disorder associated with reductions in criminal offending? A national data linkage cohort study in England. Drug and alcohol dependence, 161, 67-76.

[7] O'Keefe, C. M., & Connolly, C. J. 2010. Privacy and the use of health data for re-search. Medical Journal of Australia. 193, 9, 537-541.

[8] Brown, I., Brown, L., & Korff, D. 2010. Using NHS patient data for research without consent. Law, Innovation and Technology. 2, 2, 219-258.

[9] Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. 2004. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. International Journal of Medical Informatics. 73, 1, 1-23.

[10] De Montjoye, Yves-Alexandre, et al. 2013. Unique in the crowd: The privacy bounds of human mobility. Scientific Reports, 3.

[11] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[12] Sweeney, L. 2002. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 10, 5, 571-588.

[13] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE.

[14] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3.

[15] Armando, A., Bezzi, M., Metoui, N., & Sabetta, A. (2015). Risk-aware information disclosure. In Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance (pp. 266-276). Springer International Publishing.

[16] Eltabakh, M. Y., Padma, J., Silva, Y. N., He, P., Aref, W. G., & Bertino, E. (2012). Query processing with K-anonymity. International Journal of Data Engineering (IJDE), 3(2), 48-65.

[17] Dankar, F. K., & El Emam, K. (2012, March). The application of differential privacy to health data. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 158-166). ACM.

[18] Bertino, E., Bonatti, P. A., & Ferrari, E. 2001. TRBAC: a temporal role-based access control model. ACM Transactions on Information and System Security (TISSEC). 4, 3, 191-233.

[19] Chandran, S. M., & Joshi, J. B. 2005. LoT-RBAC: a location and time-based RBAC model. In Web Information Systems Engineering–WISE. Springer, Heidelberg, 361-375.

[20] Hansen, F., & Oleshchuk, V. 2003. SRBAC: a spatial role-based access control model for mobile systems. In Proceedings of the 7th Nordic Workshop on Secure IT Systems (NORDSEC'03). 129-141.

[21] Godik, S., Anderson, A., Parducci, B., Humenn, P., & Vajjhala, S. (2002). OASIS eXtensible access control 2 markup language (XACML) 3. Tech. rep., OASIS.

[22] Alshugran, T., & Dichter, J. (2014, May). Extracting and modeling the privacy requirements from HIPAA for healthcare applications. In Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island (pp. 1-5). IEEE.

[23] Alshugran, T., Dichter, J., & Rusu, A. (2015, May). Extending XACML to express and enforce laws and regulations privacy policies. In Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island (pp. 1-5). IEEE.

[24] McGuinness, D. L., & Van Harmelen, F. 2004. OWL web ontology language overview. W3C Recommendation. 10, 10.

[25] Knechtel, M., Hladik, J., & Dau, F. 2008. Using OWL DL reasoning to decide about authorization in RBAC. OWLED. 8, 30.

[26] Finin, T., Joshi, A., Kagal, L., Niu, J., Sandhu, R., Winsborough, W., & Thuraisingham, B. 2008. ROWLBAC: representing role based access control in OWL. In Proceedings of 13th ACM Symposium on Access Control Models and Technologies. ACM. 73-82.

[27] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. 2004. SWRL: A semantic web rule language combining OWL and RuleML. W3C Member Submission, 21, 79.

[28] O'connor, M., Knublauch, H., Tu, S., Grosof, B., Dean, M., Grosso, W., & Musen, M. 2005. Supporting rule system interoperability on the semantic web with SWRL. In The Semantic Web–ISWC 2005. Springer, Heidelberg. 974-986.

[29] Rahmouni, H. B., Solomonides, T., Mont, M. C., Shiu, S., & Rahmouni, M. (2011). A model-driven privacy compliance decision support for medical data sharing in Europe. Methods Inf Med, 50(4), 326-36.

[30] Rahmouni, H. B., Solomonides, T., Mont, M. C., & Shiu, S. (2009, August). Privacy compliance in european healthgrid domains: An ontology-based approach. In Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on (pp. 1-8). IEEE.

[31] Berners-Lee, T., Hendler, J., & Lassila, O. 2001. The semantic web. Scientific American. 284, 5, 28-37.

[32] eXtensible Access Control Markup Language (XACML) Version 3.0 OASIS Standard, 22 January 2013

[33] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). SWRL: A semantic web rule language combining OWL and RuleML. W3C Member submission, 21, 79.

[34] Xu, D., Zhang, Y., & Shen, N. 2015. Formalizing semantic differences between combining algorithms in XACML 3.0 policies. In Proceedings of IEEE International Conference on Software Quality, Reliability and Security. IEEE. 163-172.

[35] Drummond, N., & Shearer, R. (2006, October). The open world assumption. In eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web (Vol. 15).

[36] Christen, P. 2012. Privacy aspects of data matching. Data Matching. Springer, Heidelberg. 187-207.