Incorporating Personal Information into RDF

Sabah S. Al-Fedaghi, Kuwait University, P.O. Box 5969, Safat 13050, Kuwait; E-mail: sabah@eng.kuniv.edu.kw

ABSTRACT

This paper introduces an RDF-based ontology to model personal information. The triples construct is applied to 'atomic' private statements. RDF is divided into two modes: personal information mode (PIRDF) and non-personal information mode. PIRDF is a restrictive version of RDF where its constructs follow the triple format such that the semantic subject of the atomic private statement coincides with the subject of the RDF triple. Also, related (compound) pieces of personal information are made in an RDF independent collection to preserve their relationship. Literals are not allowed to embed personal information. These restrictions do not introduce any change in the syntax of RDF. Privacy protection mechanisms can be constructed upon PIRDF.

INTRODUCTION

The Resource Description Framework, RDF, is a framework for describing and interchanging machine-understandable metadata. It facilitates knowledge sharing and exchange through automated processing of Web resources in such areas as content rating, intelligent agents, intellectual property, and privacy preferences and policies. According to the World Wide Web Consortium, "Using a metadata schema to describe the formal structure of privacy practice descriptions will permit privacy practice data to be used along with other metadata in a query during resource discovery, and will permit a generic software agent to act on privacy metadata using the same techniques as used for other descriptive metadata" (Kim et al. (2002). Of special importance in the context of Semantic Web is to automate interaction regarding personal information and autonomously decide what information to exchange. A fundamental abstraction in achieving this is identifying basic 'units' of personal information.

In this paper we discuss how to define and incorporate these 'units' of personal information into a Semantic Web service model. It also proposes to construct an ontological foundation for modeling identifiable-person type of resources separate from other types of entities. Accordingly, our RDF model consists of 'person-resources' represented as nodes that refer to identifiable persons and statements about these persons. One clear advantage of such a model is that there are well-defined nodes of distinct entities: identifiable persons. In general, resources are divided into resources that represent identifiable persons and resources that identify anything else. The basic characteristic of personal information is that it uniquely identifies a real person. This person is not an interpretation that depends on namespaces. He/she is a single person who has been or was documented to exist in this world and may have several identities and descriptions. Identifiable human beings are the only "resources" that have this ontological unique identification.

Section 2 discusses related works that deal with privacy in RDF. Its contention is that there is no current work on personal information ontology below the level of specification languages for privacy policies and a user's privacy preferences. The materials in section 3 are a review of the personal information definition and classification given in Al-Fedaghi (2005), where "atomic units" (RDF triple-like constructs) of personal information are introduced. Section 4 gives our main contribution: describing refined 'atomic personal information' (called self-statements) and mapping them to (private) triples. In this paper we assume familiarity with RDF.

PRIVACY AND RDF

Our work aims at developing a model - based on personal information ontology - for RDF that is used for the purpose of analyzing and classifying personal information. 'Personal information ontology' in this paper refers to the categories of personal information that exist in the privacy domain, thus, the ontology produces a catalog that details the types of pieces of information and their relationships that are relevant for privacy (Jacob, 2003). We deal with the following problems:

(a) How to define personal information, its types, its relationships, and its mappings to its proprietors? The answers to these questions are adopted from Al-Fedaghi (2005).
(b) How to represent personal information in the RDF model? The answer to this question is accomplished through mapping personal information statements to RDF triples and imposing some restrictions on these triples.

We concentrate on a version or mode of RDF that involves only personal information. For example, *John is 32 years old,* is atomic personal information because it embeds a single referent of type *person*. *This conference is in Canada* is non-personal information because it does not embed a referent of type person. *John likes Mary* is compound personal information because it embeds two referents of type person. The atomic personal information *John's car is White* embeds two assertions:

(1) The non-personal information *The car is White,* and
(2) What we call 'self-statement': *John has a car*. This self-statement forms our 'private RDf triple' with the person-resource: John.

Currently, in such formalisms as P3P (see Al-Fedaghi (2006)) and RDF there is no explicit distinction between personal information and "owned" information. The P3P works even if the data is not "private" in the sense that it refers to a *person*. In our ontology of personal information, the system would "recognize" personal information and would distinguish it from non-personal information. Agents will be able to recognize that the requested data is private data (of the agent's owner or otherwise) and responds accordingly. We propose to mold pieces of personal information in a restricted form of RDF.

There are several attempts to link personal information to RDF. RDF is proposed as a mechanism to create a "privacy ontology group with a charter" (Hogben, 2002). Kolari et al.(2005) propose to overcome the problem of restrictive expression capabilities of APPEL through "the use of privacy policies described in an RDF based policy language, Rei, which can make policy decisions over actions and associated restrictions modeled using a Web privacy ontology" (Kolari, 2005). We claim that our ontological treatment of personal information in the context of RDF is a useful contribution to building privacy into the Semantic Web. According to Kim et al. (2002), ontology for building privacy into the Semantic Web is needed now.

PERSONAL INFORMATION

We view 'personal information' as a symbolic form that 'informs' about a single human being. 'Information' here means the 'semantic content' of a linguistic statement or assertion. A personal information theory includes a universal set of personal information agents, $Z = V \cup N$, of two fundamental types of entities: *Individual* and *Nonindividual*. *Individual* represents the set of natural persons V and *Nonindividual* represents the set of non-persons N in Z.

Definition: Personal information is any linguistic expression that has referent(s) of type *Individual*. Assuming that p(X) is a sentence such that X is the set of its referents, there are two types of personal information:

(1) p(X) is atomic personal information iff X ∩ V = is the singleton set {x }.
I.e., atomic personal information is an expression that has a single human referent.
(2) p(X) is compound personal information iff | X ∩ V | > 1.
I.e., compound personal information is an expression that has more than one human referent.

In Al-Fedaghi (2005), the relationship between individuals and their own atomic personal information is called *proprietorship*. Proprietorship of personal information is different from the concepts of possession, ownership, and copyrighting. If p is a piece of atomic personal information of v ∈ V, then p is proprietary personal information of v and v is its *proprietor*. Proprietorship gives "permanent" rights to the proprietor of personal information.

One of the most important acts on personal information is the act of *possession*. A single piece of atomic personal information may have many possessors; where its proprietor may or may not be among them. A *possessor* refers to any agent in Z that knows, stores or owns the information. Human beings are conceptualized as personal information proprietors; however, they are not the sole sources of this information. For example, companies and government agencies in N can produce and possess (non-proprietary) personal information.

Any compound personal statement is privacy-reducible to a set of atomic personal statements (Al-Fedaghi, 2005). For example *John and Mary are in love* can be privacy-reducible to *John and someone are in love* and *Someone and Mary are in love*. Atomic personal information is said to be *self-statement* if its *subject* is its proprietor and 'only its proprietor'. A framework for the derivation of self-statements from atomic personal information is given in another paper. For example, *John's house is burning* is not self-statement because it expresses two pieces of information: (a) *John has a house* and (b) *The house is burning*. Statement (a) is self-statement because its 'subject' is its proprietor. The statement (b) is non-personal statement because its 'subject' is not a person but a house. The term 'subject' here means the entity about which the information is communicated. It is an important notion when it is tied with the notion of 'subject' in RDF triples. In many cases the 'semantic subject' means that the individual affects (agent) or is-affected-by (patient) something, as reflected by the verb of the sentence. For example, in *The company invited John to an interview*, it is not clear that John is the (semantic) subject. However, *John is invited to an interview by the company* shows that John is being subjected to an action. The principle that we will follow is: the proprietor of the atomic statement has priority in being a subject when there is another entity that has a claim to being a (semantic) subject of the verb. For example in *John trained the do*g and *The dog is trained by John,* the subject is John. In *John's house is burning,* John clearly has less claim for the verb of "burning', while in *John trained the do*g, John is the trainer and the dog is the trainee, hence both have equal claim to the verb.

A self-statement's structure is the typical (subject, predicate, object) form of assertions. The (semantic) 'object' here is either an 'attribute' of the subject (e.g., tall, brave, etc.) or a non-individual entity (e.g., as in *John trained a dog*, *John derives the car*, *John loves stakes*, etc.) Every atomic personal statement is reducible to a set of self-statements and non-personal statements. Here, because of space limitation we claim that this is intuitively reasonable. It reflects the common sense notion, that a statement is about entities in reality, which can be classified into different categories (ontological objects).

PERSONAL INFORMATION RDF

The basic RDF model contains statements as parts of descriptions of some resources. We propose two modes of the RDF model:

(1) Personal Information mode of RDF (PIRDF) that facilitates all dealings with personal information.
(2) Non-personal information mode of RDF, which is the ordinary mode of RDF.

Any information in the non-personal information mode of RDF is considered as non-privacy-related information, while information in PIRDF is handled as privacy-related information of some proprietor. PIRDF may include non-personal information, but this personal information is there because it 'complements' the semantics of some personal information. For example, in *John takes drug x23 which is used to treat cancer*, both the self-statement *John takes a drug* and the non-personal information *The drug is called x23 and used to treat cancer* are handled by PIRDF. These statements are treated as a collection of triples in order to facilitate reconstructing the semantics of the original statement, if such reconstruction is needed. Also, treating a set of statements as an RDF collection is the method used in PIRDF to represent compound personal information.

Personal information in PIRDF is treated in a special way to protect the privacy of proprietors. We will not deal with this side of PIRDF, and concentrate in this paper on the method of describing personal information in PIRDF. Our strategy is not to introduce any new feature to the standard RDF, rather we propose measures that restrict modeling triples in a way suitable for personal information. These restrictions are as follows:

(1) The subject of any RDF triple that represents self-statement is always a proprietor. It is also allowed to have the proprietor as the object in a reification statement with attribute: subject. The reason for making the proprietor as the subject of this type of triples is because it is the "semantic" subject of the corresponding self-statements. Thus, we merge the 'triple subject' with the semantic subject. This provides unique identification of personal information according to the content of the triple (subject + type: person), hence, it is not necessary to 'RDF-type' private statements. The resulting graph reflects the proprietor as the center of his/her personal information sphere as shown in figure 1.

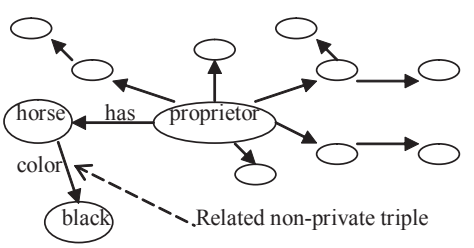
Compound personal information is represented as a set of triples of atomic personal information in an RDF collection. The reason for this is to allow the system to clearly identify the proprietor of any piece of personal information. Thus, the triple (John, threatens, Alice) is not allowed and replaced by (John, threatens, someone) and (Alice, being-threatened-by, someone). The method of representing these two triples without persons being allowed as objects is an open problem. Notice that the purpose in this type of modeling is to clearly distinguish John's personal information from Alice's personal information. It is possible that Alice is not permitted to know who threatens her (e.g., mental health confidentiality); however, she may have the right to know that she is the object of a threat.

A "resource" is defined in RDF as *anything that has identity*. In general, according to RFC 2396 (Berners-Lee et al., 1998) "A resource can be anything that has identity... Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources." PIRDF describes a special type of resources: 'identified persons' class that is defined as the RDF schema vocabulary:

:Proprietors rdf:type rdfs:Class.

We denote this type of resource as 'person resource', for short, *pe*source. This pesource is an *information entity* that is realized through a set of atomic personal information. Every pesource is uniquely identified in the PIRDF. A person as an 'information entity' is a known ontological concept. According to Floridi (1998), all objects including human beings are 'information objects': "[A] person, a free and responsible agent, is after all a packet of information... We are our information... personal information is a constitutive part of a me-hood" (Stein et al., 2000). We observe that there is a difference between the conceptualization of a human being as an information entity and as a personal information entity. Thus, for us a pesource is an information entity formed from a set of pieces of atomic personal information.

Figure 1. The proprietor is always the (semantic and syntactical) subject of the private triples



PIRDF maintains the informational ontology of each proprietor through maintaining his/her personal information. It identifies pieces of personal information of each proprietor. These pieces of personal information are treated with special consideration in terms of: operations that include: disclosure, possession, consistency, sharing, etc. Non-personal information is treated in an ordinary way.

Using our terminology for personal statements, we can categorize triples that correspond to personal statements as follows:

(a) A non-private triple is a triple that does not contain any resource that denotes a person.
(b) A private triple is a triple that represents a self-statement.

In RDF, the subject is the node the statement is about. It is either an URI reference or a blank node. In PIRDF, the proprietor is always the subject of private triple. We assume that all personal information statements are in the form of self-statement. The subject of personal information triple refers to the subject of a self-statement.

Example: Suppose that we have the atomic information *Ernest Hemingway's FAREWELL TO ARMS is located at AF.123*. It embeds the two statements:

(a) Ernest Hemingway authored FAREWELL TO ARMS
(b) FAREWELL TO ARMS is located at AF.123

PIRDF gives special considerations to represent facts about persons, not about books, houses, etc. Assume that Ernest Hemingway is an entity of type person declared as proprietor in :Proprietors**.** Being a subject of triple (a) represents an implied type of the triple, thus we don't have to introduce the type 'private' to be associated with statement (a). A private triple is a triple where the subject is of type person. This is a very useful convention.

Alternatively, We could have considered both:

Ernest Hemingway authored FAREWELL TO ARMS,
FAREWELL TO ARMS is authored by Ernest Hemingway

as forms of personal information. However, in this case we have to distinguish between the "semantic subject" (what the statement is about) and the RDF subject. In *FAREWELL TO ARMS is authored by Ernest Hemingway,* the (privacy) "semantic object" is Ernest Hemingway while the RDF subject is (the book): FAREWELL TO ARMS.

In the original example, the other piece of information: *FAREWELL TO ARMS is located on AF.123* is non-personal information. Hence, it has no special consideration in PIRDF. It could be represented, if we like, as *AF.123 is the location of FAREWELL TO ARMS*.

In PIRDF world, the distinction between private and non-private triples is important. PIRDF is a partial, simplified conceptualization of the world created for the purpose of handling personal information and defined in a formal, machine-processable language.

In RDF, URIs identify network-accessible things, things that are not network-accessible, and abstract concepts. In PIRDF, identifiable network-accessible 'things' are of two kinds: privacy-based and non-private-based things. Without loss of generality, we assume privacy-based things are textual materials. Similarly, things that are not network-accessible are categorized into persons and non-persons.

In RDF, a URI is just a node that has a URI label on it, where the URI identifies the resource represented by the node. Since the URI directly identifies the resource represented by a node, RDF assumes that nodes with the same URI represent the same resource. A URI may be complemented with an optional fragment identifier, URIref. In PIRDF each proprietor is mapped uniquely to a single entry in the vocabulary pirdf:proprietor. This does not prevent from using different synonyms that are mapped to a single URIref in pirdf:proprietor. In principle, a proprietor may choose to have several "personal information personalities" through having more than one URIref in pirdf:proprietor. In this case he/she has two different "personal information spheres." This is an implementation issue similar to the problem of the uniqueness of RDF resources. However, the uniqueness of proprietors is easier to handle because of the already ultra-importance of identification of persons inside and outside the network. We will assume that each proprietor identifier represents a single person.

Example: Consider the following familiar compound personal information:

Ralph Swick says that Ora Lassila is the creator of the resource http://www.w3.org/Home/Lassila.

According to the W3C Recommendation (1999), figure 2 represents its graph form. We can criticize this graph representation on the ground that it does not correspond with the linguistic structure of the statement, which is in the triple form,

(*Ralph Swick says* (*Ora Lassila is the creator of the resource http://www.w3.org/Home/Lassila*)).

The main "subject" in the graph is the 'statement', while the original main subject *Ralph Swick* has become a value of an attribute to the statement. So, semantically, the whole graph is about 'the statement': about its subject, its object, its predicate, and its 'attributer'. This is not a suitable graph representation in PIRDF, because personal information is always about the proprietor: he/she is the subject.

Figure 3 shows the graph form of the given statement in PIRDF.

It has two atomic statements:

(1) *Ora Lassila is the creator of the resource http://www.w3.org/Home/Lassila.*
(2) Ralph Swick says statement (1)

This example illustrates another allowable position of the proprietor in PIRDF: as an object of the attribute: subject. Clearly, this case is syntactically discoverable.

The two atomic pieces of personal information (1) and (2) embed identities of proprietors. The personal information (1) is represented in the shown triples. Notice that PIRDF assumes that the only way to identify a proprietor is through his/her pirdf:proprietor identification.

Example: Suppose the we want to express the statement (Johannesen, 2004): *Tom borrowed a book from Mike*. In RDF, this can be expressed through a blank node that has connections to different properties as follows:

Figure 2. Reification in RDF

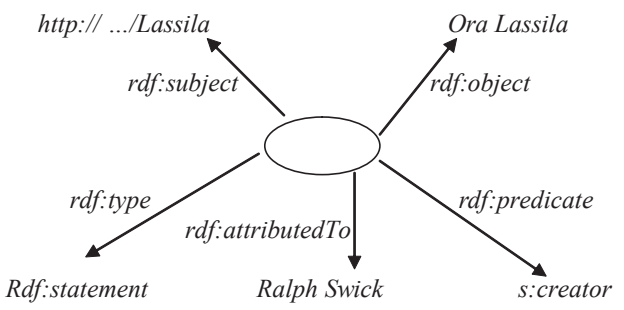
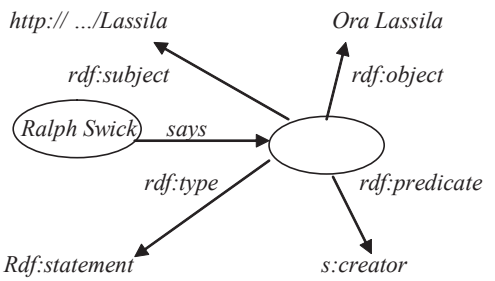


Figure 3. Reification in PIRDF



_:a rdf:type x:Borrowing
_:a x:who prs:Tom
_:a x:fromWhom prs:Mike
_:a x:what _:b
_:b rdf:type x:Book

This specification is described as "not the easy way to do the Semantic Web" (Johannesen, 2004). In PIRDF, the compound personal information *Tom borrowed a book from Mike* can be implemented as a collection of two triples: *Tom borrowed a book* and *Mike has a book* as shown in figure 4. Thus, the notion of compound personal information embeds the concept of a set.

Example: Consider the familiar RDF expression: (:Jane :daughterOf :John, : Jennifer). It can be represented in PIRDF as shown in figure 5.

Different personal assertions are distinguished as follows:

The personal information of Jane: *Jane is the daughter of some parents*
The personal information of John: *John is the father of someone*
The personal information of Jennifer: *Jennifer is the mother of someone*

It may sound somewhat odd to say that *Jane is the daughter of some parents*. However, imagine that the graph represents the database of an adopting agency. Then even Jane should not know her parents without permission.

The design of PIRDF requires several modifications with regard to literals. The general rule here is, it is not allowable to make literal personal information. Thus, *The newspaper headline is: John is a killer* may be described in RDF as the triple (<http://…newspaper> <http://…headline> "John is a killer"). However, in PIRDF, it is necessary to factor out personal information as shown in figure 6.

CONCLUSION

We have introduced elements of 'personal information modeling' in RDF. The proposed model is based on two foundations: defining personal information in terms of statements that refer to persons, and representing statements as RDF triples. The result is a preliminary RDF-based ontology of personal information. Space limitation does not allow more details to achieve further specification of such ontology. Our contribution is a first step towards focusing on the problem of personal information ontology below the level of modeling privacy preferences and policies.

Figure 4. Graph of Tom borrowed a book from Mike in PIRDF

Figure 5. Jane is the daughter of John and Jennifer in PIRDF

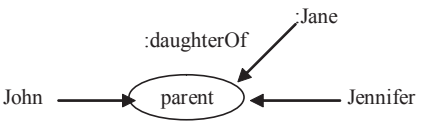
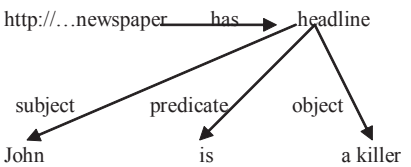


Figure 6. PIRDF allows the proprietor to be the object, if the attribute is 'subject'



There are several extensions to the basic PIEDF model. For users, their triples include:

- Private triples and related non-private triples,
- Sets of triples that represent compound personal information,
- Triples that represent personal information in the possession of the user.

Thus, in building a rules system, each of these types of personal information is treated differently. Generally, how PIRDF influences the rule system and policy language needs to be investigated. Also, several constructs of PIRDF present interesting issues in the new formalisms such as OWL (see Al-Fedaghi (2006)).

REFERENCES

Al-Fedaghi, S. (2006). Personal Information Flow Model for P3P, W3C Workshop on Languages for Privacy Policy Negotiation and Semantics-Driven Enforcement, Ispra (Italy), 17-18.

Al-Fedaghi S. and Ahmad, M. (2006). Personal Information Modeling in Semantic Web, The Asian Semantic Web Conference (ASWC), Beijing, China, September 3-7.

Al-Fedaghi, S. (2005). How to Calculate the Information Privacy. Proceedings of the Third Annual Conference on Privacy, Security and Trust, October 12-14, St. Andrews , New Brunswick, Canada.

Berners-Lee T., Fielding, R. Irvine, U. C. and Masinter, L. (1998). RFC 2396: Uniform Resource Identifiers (URI): Generic Syntax, August. http://www.ietf.org/rfc/rfc2396.txt?number=2396

Floridi, L. (1998). Information Ethics: On the Philosophical Foundation of Computer Ethics, ETHICOMP98 The Fourth International Conference on Ethical Issues of Information Technology. http://www.wolfson.ox.ac.uk/~floridi/ie.htm.

Hogben, G. (2002). Development of a Data Protection Ontology, Joint Research Centre, Ispra, 27 May. http://64.233.179.104/search?q=cache:tMF0U4WW_iQJ:p3p.jrc.it/presentations/Data%2520Protection%2520Ontology.v2.ppt+privacy+rdf&hl=en

Jacob, Elin K. (2003). Ontologies and the Semantic Web, Bulletin of the American Society for Information Science & Technology, Apr/May. http://www.findarticles.com/p/articles/mi_qa3991/is_200304/ai_n9235530

Johannesen A. (2004). shelter.nu, 24 Mar 2004, http://www.shelter.nu/blog-078.html

Kim, A., Hoffman, L. J., and Martin, C. D. (2002). Building Privacy into the Semantic Web: An Ontology Needed Now, Semantic Web Workshop, Hawaii USA, http://semanticweb2002.aifb.uni-karlsruhe.de/proceedings/Position/kim2.pdf

Kolari P., Li D., Shashidhara G. Joshi, A. Finin, T. and Kagal, L. (2005). Enhancing Web Privacy Protection through Declarative Policies, proceedings of the IEEE Workshop on Policy for Distributed Systems and Networks (POLICY 2005), June, 2005. http://ebiquity.umbc.edu/_file_directory_/papers/156.pdf

Resource Description Framework (RDF) Schema Specification (1999). W3C Proposed Recommendation 03 March, http://www.w3.org/TR/1999/PR-rdf-schema-19990303/

Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999

Stein, L., D. Connolly, D. McGuinness (2000). DAML-ONT Initial Release, http://www.daml.org/2000/10/daml-ont.html