Towards the Anonymisation of RDF Data

Filip Radulovic
Ontology Engineering Group
ETSI Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
fradulovic@fi.upm.es

Raúl García-Castro
Ontology Engineering Group
ETSI Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
rgarcia@fi.upm.es

Asunción Gómez-Pérez
Ontology Engineering Group
ETSI Informáticos
Universidad Politécnica de Madrid
Madrid, Spain
asun@fi.upm.es

***Abstract*—Privacy protection in published data sets is of crucial importance, and anonymisation is one well-known technique for privacy protection that has been successfully used in practice. However, existing anonymisation frameworks have in mind specific data structures (i.e., tabular data) and, because of this, these frameworks are difficult to apply in the case of RDF data. This paper presents an RDF anonymisation framework that has been developed to address the particularities of the RDF specification. Such framework includes an anonymisation model for RDF data, a set of anonymisation operations for the implementation of such model, and a metric for measuring precision and distortion of anonymised RDF data. Furthermore, this paper presents a use case of the proposed RDF anonymisation framework.**

I. Introduction

Large quantitiesof data are gathered and published every day by public and private companies and institutions. One key aspect of data publishing is the protection of the privacy of entities of interest (e.g., individuals), and failure to ensure the privacy can not only harm the reputation of a publisher, but can also compromise the privacy of entities of interest by making their private information available to third parties.

Ensuring the privacy of data while preserving data usefulness is not a simple task. Usually, removal of the data that explicitly identify the entity of interest, such as social security numbers or telephone numbers, does not alone ensure privacy since the remaining data can often be linked to other published data and used for identification purposes [1]. For example, Sweeney showed in an experiment that 87% of the U.S. population is likely to be uniquely identified based only on a combination of a ZIP code, gender, and date of birth [2].

Anonymisation is one technique for privacy protection that has been successfully applied in practice, and a number of anonymisation frameworks have been developed to this date. However, these frameworks are developed having in mind specific data structures, such as tabular data, and they are difficult to apply for the anonymisation of data that have different structures and formats. This is the case of RDF (Resource Description Framework) [3] data.

With the increasing amount of RDF data being published in the Web (usually as Linked Data), privacy issues are expected to emerge. Furthermore, privacy concerns hinder the publication of Linked Data in different sectors (e.g., healthcare, energy) where the re-identification of individuals or other entities of interest can lead to social or legal issues.

DOI reference number: 10.18293/SEKE2015-167

This paper presents a framework for the anonymisation of RDF data. Such framework describes an anonymisation model for RDF data called *k-RDFanonymity*, as well as anonymisation operations for the implementation of the mentioned model and a metric for measuring the precision and distortion of anonymised RDF data. Furthermore, this paper also presents a use case of the presented RDF anonymisation framework.

The reminder of this paper is organised as follows. Section II describes related work, while Section III presents the framework for the anonymisation of RDF data. Section IV presents a use case of such framework and, finally, Section V draws some conclusions and includes ideas for future work.

II. Related Work and Background

This section gives a brief description of related work. First, we describe the foundations of data anonymisation frameworks. Afterwards, we describe RDF and its particularities that are of interest for data anonymisation.

A. Data Anonymisation Frameworks

Anonymisation is a widely accepted and used framework for privacy-preserving data publishing that aims to ensure the privacy of data and balance data analysis and utility [4].

In privacy-preserving data publishing, there are several data attributes of the entity of interest that are taken into account:

- *Explicit identifiers* are attributes that explicitly identify the entity of interest (e.g., identifier of a person, property number of a building).
- *Quasi identifiers (QIDs)* are sets of attributes that can potentially identify the entity of interest. Usually, those are the attributes whose values can be found in other data sets and that can be then used for identification purposes (it is important to note that each attribute in a quasi identifier does not alone identify an entity of interest).
- *Sensitive attributes* are those attributes that describe some sensitive information about the entity of interest (e.g., salary, disease).
- *Non-sensitive attributes* are all attributes which do not belong to any of the previous categories.

Explicit identifiers, QIDs, and sensitive attributes can be considered to be private attributes of the entity of interest.

Non-sensitive attributes are not considered to be a privacy issue.

To illustrate these attributes, we present an example of a medical records with data about patients and their diseases (Table I). There are no definitive guidelines on how to properly classify attributes, so the classification can be a difficult task. The attributes present in our example data set can be classified as follows: *Id* is an explicit identifier since it can explicitly identify the patient that a record belongs to; a set of attributes *{Job, Age}* is a quasi identifier (QID) since it can be expected that the same set of attributes can appear in some other data set that can also contain additional data (but not diseases) that are sufficient for patient identification (e.g., name and surename); *Disease* is a sensitive attribute since it gives sensitive information about patients.

TABLE I: Example of patients' medical records data.

Id	Job	Age	Disease
1872	Teacher	24	HIV
1352	Lawyer	28	Flu
1453	Musician	32	Flu
1389	Writer	35	HIV
1463	Writer	36	HIV
1305	Lawyer	22	Flu
1435	Teacher	25	HIV
1058	Musician	38	Flu

Data anonymisation implies that explicit identifiers must be removed from the data set [4] and that the original QIDs are anonymised. Different anonymisation models have been developed having in mind tabular data structures (e.g., k-anonymity [5], l-diversity [6]), and these anonymisation models can be implemented by applying various anonymisation operations [4]:

- *Suppression* is a technique in which one or more values in a data set are removed or replaced with some special value, while removed or replaced values are not disclosed.
- *Generalisation* is a technique that transforms values into more general values, i,e., into new values that are less precise but still consistent with the original ones.
- *Anatomisation* implies that the relationship between the quasi identifiers and the sensitive values is removed, while the data is not modified. This is achieved by separating the data related to quasi identifiers from the data containing sensitive values and by providing the relationship between the two data sets by introducing an identifier.
- *Perturbation* is a technique in which the original data are replaced with noise or synthetic data in such a way that statistical analyses based on the perturbed data do not significantly differ from the statistical analysis of the original data [4]. Unlike previous techniques, perturbation does not preserve the truthfulness of the data and the perturbed data do not correspond to real world entities.

Anonymisation models and anonymisation operations are integral parts of a data anonymisation framework. By applying anonymisation operations, an anonymisation model is implemented and thus, the privacy of entities of interest in a data set is ensured. Furthermore, a data anonymisation framework specifies various metrics for measuring the distortion and usefulness of anonymised data (e.g., precision and minimal distortion [7], [8]).

B. Anonymisation in Resource Description Framework

RDF is a specification for describing resources on the Web, where resources can be anything including documents, objects, people, or abstract concepts [3]. Unlike in tabular data formats (e.g., databases), where existing anonymisation frameworks have been successfully applied to this date, data in RDF are structured in a different manner as a graph.

The key concept in RDF is an *RDF statement (s,p,o)*, also called an *RDF triple*, which consists of a subject, a predicate, and an object, and which encodes a claim about the world. Each RDF triple implies the existence of a relationship that holds between two resources denoted by a subject and an object. Relationships in RDF triples are denoted by a predicate, also called a property, and are directed from subject to object.

While in tabular data formats attribute values of entities of interest are stored in columns, attribute values of entities of interest described in RDF appear as resources, either as IRIs or as literals, that describe these entities of interest. In an RDF graph, literals appear only as objects, while IRIs can appear in several places: as subjects, as predicates, or as objects. Therefore, attribute values of entities of interest can have different forms and can appear in different places in the RDF descriptions of these entities of interest.

In an RDF graph, resources can be classified according to categories specified by classes that belong to a specific *vocabulary*. Vocabularies are used in combination with RDF for providing semantic information about resources, and are defined using a specific language (e.g., RDF Schema or OWL). Relationships between an RDF resource and its class are defined through an *rdf:type* property. These relationships also describe resources in RDF.

The particularities of the RDF model imply difficulties in the direct application of existing anonymisation frameworks, which were developed having in mind different data structures and formats. Therefore, in order to successfully anonymise RDF data, anonymisation frameworks that address the particularities of RDF are needed. Although some effort in this direction exists [9], it addresses only generalisation and suppression, and it does not include any anonymisation metrics.

III. A Framework for the Anonymisation of RDF Data

This section describes a framework for the anonymisation of RDF data. Such framework is based on existing anonymisation frameworks and has been specifically defined to take into account the particularities of the RDF specification. It consists of an anonymisation model, of a set of anonymisation operations, and of an anonymisation metric adapted to fit the RDF specification.

A. Privacy-related Entity Attributes in RDF

One characteristic of RDF, which is of relevance for the problem of anonymisation, is that entity attribute values can appear in different places and forms in the description of resources that represent entities of interest. Because of this, privacy-related attributes in RDF are more difficult to identify and addressing the privacy of RDF data can be a complex task.

In the RDF description of entities of interest, values of privacy-related entity attributes (explicit identifiers, quasi identifiers, and sensitive attributes) can appear in:

- *Resource IRI*. Values for all three types of attributes can appear in the IRI of a resource, exposing them to humans.
- *Datatype property value*. Values for all three types of attributes can appear as literals in datatype property values in a resource description (object in a statement).
- *Object property value*. Values for all three types of attributes can appear as IRIs in object property values in a resource description (object in a statement).
- *Property IRI*. Values for attributes that belong to quasi identifiers and for sensitive attributes can appear in the IRI of a property in a resource description. Having explicit identifiers appear as property IRIs is not expected.
- *Related resource*. Values for all three types of attributes can also appear in more complex scenarios in the description of resources that are related to resources that represent entities of interest through RDF properties. In this case, values can appear in all scenarios presented above: in resource IRI, datatype property value, object property value, property IRI, or another related resource.

Next, we present an example of the previous scenarios[1] (Listing 1) in which the first record from Table I is described in RDF using the Turtle syntax. In this example, the identifier appears in the resource IRI, the age appears as a datatype property value, the job appears as an object property value, and the disease appears as a property IRI.

```
1 <http://example.com/resource/Person/1872> a foaf:Person;
2     foaf:age "24"; ex:hasJob ex:Teacher; ex:hasHIV "true".
```

Listing 1: Example of a patient's medical record in RDF.

B. RDF Anonymisation Model

This section presents an anonymisation model for RDF, called *k-RDFanonymity*. This model is inspired by the k-anonymity model [1], [5] developed by Samarati and Sweeney for the anonymisation of tabular data.

Definition 1. A subgraph G_r of an RDF graph G ($G_r \in G$) describes an entity of interest represented by a resource r if G_r is the union of all the subgraphs of G that include information about attributes that describe the entity of interest represented by r, regardless of whether r is the subject or the object of statements in these subgraphs.

Definition 2 (Equivalence). Graphs G_{r_1} and G_{r_2} are equivalent ($G_{r_1} \equiv G_{r_2}$) $\Leftrightarrow \forall$(s$\neq r_1$,p,o$\neq r_1$) $\in G_{r_1}$ $\exists$(s$\neq r_2$,p,o$\neq r_2$) $\in G_{r_2} \wedge \forall(r_1,p,o) \in G_{r_1}$ $\exists(r_2,p,o) \in G_{r_2} \wedge \forall(s,p,r_1) \in G_{r_1}$ $\exists(s,p,r_2) \in G_{r_2}$.

Definition 3. In an RDF graph G, with *QID(G)* we denote a set of QID attributes that describe any entity of interest represented by a resource in G.

Definition 4. A subgraph G_r of an RDF graph G describes an entity of interest represented by a resource r with respect to *QID(G)*, written $G_r(QID(G))$, if G_r includes information about all QID attributes.

Definition 5 (k-RDFanonymity). Let I be a set of resources that represent entities of interest described in an RDF graph G, and let *QID(G)* be a set of QID attributes that describe these entities of interest. kRDF-anonymity in G is satisfied $\Leftrightarrow \forall G_r(QID(G)) \in G$, $r \in I$, $\exists r_s \in I$, $s \in$ [1,k-1] $\Rightarrow \forall G_{r_s}(QID(G)) \in G$, $G_{r_s} \equiv G_r$. An RDF graph that satisfies this premised is called k-RDFanonymous.

In a k-RDFanonymous graph, each resource r that represents an entity of interest cannot be distinguished from k-1 other resources that represent entities of interest in a graph with respect to *QID(G)*. Therefore, the probability of identifying a specific resource based on resource descriptions with respect to *QID(G)* is 1/k.

C. RDF Anonymisation Operations

This section presents different anonymisation operations that can be used for implementing the k-RDFanonymity model. These operations are based on the anonymisation operations presented in Section II-A and address private attributes, i.e., explicit identifiers, quasi identifiers, and sensitive attributes.

1) Generalisation and Suppression: In the context of RDF, suppression denotes that a resource (i.e., an IRI or literal) is completely removed or replaced with some specific resource while the original replaced resource is not disclosed in any way. Generalisation denotes that the original resource is replaced with other resource that describes a more general concept.

The starting point of generalisation and suppression is a domain generalisation hierarchy of a an attribute, in which the elements of the hierarchy are resources that represent attribute values (i.e., that include information about that attribute). Figure 1 shows generalisation hierarchies for the *Age* attribute (literals), the *Job* attribute (IRIs representing a class from a specific vocabulary), and the *Disease* attribute (properties).

While generalisation implies the use of more general resources from a hierarchy (e.g., *ex:Job* instead of *ex:Musician*), suppression implies the complete removal of a resource or the use of the resource at the top of the hierarchy (e.g., *owl:Thing* instead of *ex:Musician*).

Since resources that include information about explicit identifiers unequivocally identify entities of interest, these resources have to be suppressed, while resources that include information about QID and sensitive attributes can be generalised or suppressed, depending on the concrete scenario. Next, we describe generalisation and suppression through different scenarios depending on the position in which resources that include information about explicit identifiers, QID and sensitive

[1]In the sake of simplicity, we omit prefix and datatype declarations in all the examples.

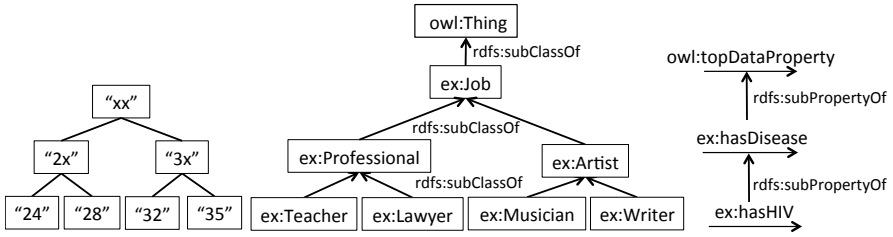

Fig. 1: Generalisation hierarchies for Age (left), Job (center) and Disease (right).

attributes appear in the description of a resource that represents an entity of interest (Section III-A):

- *Resource IRI*. If a resource that includes information about an explicit identifier appears in the IRI of the resource that represents an entity of interest, it cannot be simply removed from the graph because it would result in the loss of connections between nodes. Therefore, it has to be suppressed by replacing the original IRI with another one. This can be achieved through: i) simple replacement of the IRI with some arbitrary IRI. In this case, the uniqueness of each IRI has to be ensured; ii) encoding of the IRI by using encryption or hashing. In this case, the encryption or hashing function should ensure that the original IRI cannot be easily recovered, and that new IRI is unique; or iii) replacing the IRI node with a blank node. In this case, the original resource that includes information about the explicit identifier is suppressed and the original IRI is not disclosed.
  Resources that include information about QID and sensitive attributes that appear in the IRIs of resources that represent entities of interest can be addressed by generalising or suppressing the original IRIs, using the previously defined domain generalisation hierarchies.
  It is important to note that generalisation and suppression of IRIs breaks the uniqueness of the IRIs in a graph. This can be solved by introducing additional unique identifier values as part of the IRI of each resource.
- *Property values*. Resources that include information about explicit identifiers that appear either as datatype or as object property values are addressed by complete suppression, i.e., by removing property values from the graph. Resources that include information about QID and sensitive attributes that appear either as datatype or as object property values can be addressed by generalising or suppressing the original resources that appear as property values. In some cases, this operation has to be supported by the vocabulary design.
- *Property IRI*. Resources that include information about QID and sensitive attributes that appear as property IRIs can be addressed by generalisation through the use of super-properties.

Listing 2 presents an example of applying generalisation and suppression on the RDF data presented in Listing 1. The resource that includes information about the explicit identifier is addressed by suppressing the original IRI with another one. Age (which is described as a datatype property value and includes information about a QID attribute), job (which is described as an object property value and also includes information about a QID attribute), and disease (which is described as a property IRI and includes information about a sensitive attribute) are addressed by generalising by one level in the generalisation hierarchy.

```
1 <http://example.com/resource/Person/Per01> a foaf:Person;
2     foaf:age "2x"; ex:hasJob ex:Professional;
3     ex:hasDisease "true".
```

Listing 2: Example of generalisation and suppression in RDF.

Since generalisation reduces the semantic precision of information and since for an RDF graph G there can be multiple generalisations, different generalisations of an RDF graph G are characterised by different semantic precisions of information.

Definition 7 (Graph generalisation). An RDF graph G_g is a generalisation of an RDF graph G with respect to $QID(G)$, written $G_g(QID(G)) \geq G(QID(G))$, if some or all resources that include information about QID attributes in a graph G are generalised.

Definition 8 (kRDF-minimal generalisation). Let an RDF graph G_g be a generalisation of an RDF graph G with respect to $QID(G)$. $G_g(QID(G))$ is said to be the k-RDFminimal generalisation of an RDF graph G with respect to $QID(G) \Leftrightarrow G_g(QID(G))$ is a k-RDFanonymous graph $\wedge$ $\forall G_i(QID(G))$: $G_i(QID(G)) \geq G(QID(G))$, $G_g(QID(G)) \geq G_i(QID(G))$, $G_i(QID(G))$ is a k-RDFanonymous graph $\Rightarrow$ $G_g(QID(G)) \equiv G_i(QID(G))$.

2) Anatomisation: In the context of RDF, anatomisation implies that resources that include information about sensitive attributes are not directly connected to resources r_i that represent entities of interest. Instead, resources that include information about sensitive attributes are grouped into several groups which describe how many resources r_i belong to each group. All resources r_i are then connected to these groups. This way, for each resource r_i it is known to which group it belongs to and, hence, it is only known with how many other resources that represent entities of interest r_i shares the sensitive information from a particular group.

Listing 3 shows an example of applying anatomisation on the RDF data presented in Listing 1.

From the previously described graph, for any disease group there is only information about how many patients have each

disease. Therefore, for any resource (patient) it is not explicitly known which disease is associated with it.

```
<http://example.com/resource/Person/Per01> a foaf:Person;
    foaf:age "24"; ex:hasJob ex:Teacher;
    ex:inGroup <http://example.com/resource/Group/01>.
<http://example.com/resource/Group/01> a ex:DiseaseGroup;
    ex:hasDisease <http://example.com/resource/Disease/X>;
    ex:hasDisease <http://example.com/resource/Disease/Y>.
<http://example.com/resource/Disease/X> a ex:Disease;
    ex:name "HIV"; ex:cardinality "4".
<http://example.com/resource/Disease/Y> a ex:Disease;
    ex:name "Flu"; ex:cardinality "4".
```

Listing 3: Example of anatomisation in RDF.

3) Perturbation: In the context of RDF, perturbation is an operation which replaces original resources in such a way that the semantics of resources affected is also changed, while preserving the statistical information of the original RDF graph. This can be achieved by: i) adding noise to the RDF data in such a way that the semantics of data is changed; ii) swapping resources by assigning to a resource that represents an entity of interest the description of some other resource that represents another entity of interest; or iii) generating synthetic resources.

Listing 4 presents an example of applying perturbation on the RDF data presented in Listing 1. In this example, noise has been introduced to the resource describing age, while resources describing job and disease have been swapped.

```
<http://example.com/resource/Person/Per01> a foaf:Person;
    foaf:age "22"; ex:hasJob ex:Musician; ex:hasFlu "true".
```

Listing 4: Example of perturbation in RDF.

D. RDF Information Metrics

In the situation where for a given RDF graph there exist multiple k-RDFanonymous graphs, the decision on which k-RDFanonymous graph is the best to use for privacy protection can be a difficult task. In order to provide information that can help in making this decision, we have defined *RDFprec*, a precision metric of a k-RDFanonymous graph. This metric is based on the precision metric for tabular data developed by Sweeney [8] and can be used for defining the minimal distortion of an RDF graph.

Definition 9. With DGH_{a_i} we denote a domain generalisation hierarchy of an attribute a_i which describes an entity of interest represented by a resource. With $|DGH_{a_i}|$ we denote the number of levels in DGH_{a_i}, where the lowest level in a hierarchy is level *0*. With v_{ij} we denote a resource which includes information about the value of an attribute a_i, and which describes an entity of interest represented by a resource r_j. With $h(v_{ij})$ we denote a height of a resource v_{ij} in DGH_{a_i}, where a resource v_{ij} at the lowest level in the hierarchy has height *0*. With $|r|$ we denote the number of resources r that represent entities of interest in a graph G, i.e., those resources that have to be anonymised.

An example of a domain generalisation hierarchy is shown on Figure 1. In this example, in the case of a resource that describes age there are 2 levels, while in the case of a resource that describes jobs there are 3 levels in the domain generalisation hierarchy. Resource "*2x*" is on the first level in the hierarchy.

Definition 10 (RDFprec). Let G be an RDF graph, G_g be a generalisation of G, v_{ij} be a generalised resource from G_g which includes information about the value of an attribute a_i and which describes an entity of interest represented by a resource r_j, DGH_{a_i} be the domain generalisation hierarchy of an attribute a_i, n_a be the number of generalised attributes, and m_i be the number of entities of interest represented by resources r_j in which a resource that includes information about an attribute a_i is generalised. The precision of G_g, written *RDFprec(G_g)* is defined with the following formula:

$$\text{RDFprec}(G_g) = 1 - \frac{\sum_{i=1}^{n_a} \sum_{j=1}^{m_i} \frac{h(v_{ij})}{|DGH_{a_i}|}}{|r| * n_a}$$

If all resources in an RDF graph G_g are generalised to the highest level in a domain generalisation hierarchy, each $h(v_{ij}) = |DGH_{a_i}|$ and RDFprec(G_g) = 0. Contrary, if there are no resources that are generalised, each $h(v_{ij}) = 0$ and the RDFprec(G_g) = 1.

As an example, precision of the example RDF graph G presented in Listing 2, which consists of only one resource that represents an entity of interest, with respect to the domain generalisation hierarchy on Figure 1 is

$$\text{RDFprec}(G) = 1 - \frac{1/2 + 1/3 + 1/2}{1 * 3} = \frac{5}{9}$$

Definition 11 - RDF minimal distortion. Let an RDF graph G_g be a generalisation of an RDF graph G with respect to *QID(G)*. *G_g(QID(G))* is said to be the RDFminimal distortion of an RDF graph G with respect to *QID(G)* $\Leftrightarrow$ *G_g(QID(G))* is a k-RDFanonymous graph $\wedge$ $\forall$*G_i(QID(G))*: RDFprec(G) $\geq$ RDFprec(G_i), RDFprec(G_i) $\geq$ RDFprec(G_g), *G_i(QID(G))* is a k-RDFanonymous graph $\Rightarrow$ *G_g(QID(G))* $\equiv$ *G_i(QID(G))*.

IV. Use of the RDF Anonymisation Framework

This section presents a use case of the RDF anonymisation framework presented in this paper. In this use case, we examine different generalisations of an RDF graph G that describes the medical records presented in Table I, based on the domain generalisation hierarchies presented in Figure 1.

For graph G six different suppressions and five different generalisations are possible. In this example, we focus only on generalisations that are addressed through a generalisation of the QID attributes, which include *Job* and *Age*.

The original RDF graph G with no generalisation, written $G_{[0,0]}$, describes the values for the *Job* and *Age* attributes at the lowest (zero) level in the domain generalisation hierarchies, and corresponds to the data in Table I. In this case, the k constraint is *0* and RDFprec($G_{[0,0]}$) = 1. Listing 5 shows the excerpt of the original graph $G_{[0,0]}$ which includes information about teachers and lawyers.

Possible generalisations of graph G are: $G_{[0,1]}$ (meaning *Job* is not generalised while *Age* is generalised by one level), $G_{[1,0]}$, $G_{[1,1]}$, $G_{[2,0]}$, and $G_{[2,1]}$. Graphs $G_{[0,1]}$, $G_{[1,1]}$, and $G_{[2,1]}$ are k-RDFanonymous for k=*2*, while $G_{[1,0]}$ and $G_{[2,0]}$ are not k-RDFanonymous since in these cases k = *0*.

```
<http://example.com/resource/Person/Per01> a foaf:Person;
    foaf:age "24"; ex:hasJob ex:Teacher; ex:hasHIV "true".
<http://example.com/resource/Person/Per02> a foaf:Person;
    foaf:age "28"; ex:hasJob ex:Lawyer; ex:hasFlu "true".
<http://example.com/resource/Person/Per06> a foaf:Person;
    foaf:age "22"; ex:hasJob ex:Lawyer; ex:hasFlu "true".
<http://example.com/resource/Person/Per07> a foaf:Person;
    foaf:age "25"; ex:hasJob ex:Teacher; ex:hasHIV "true".
```

Listing 5: Excerpt of the original graph with no generalisation.

Listing 6 presents the excerpt of the generalisation graph $G_{[0,1]}$ related to the excerpt presented in Listing 5. We can observe that, for any given combination of resources that include information about the *Job* and *Age* attributes, there are two resources that represent patients that are described with that same combination. For example, *Per01* and *Per07* are both teachers that are between 20 and 30 years old, and they both have HIV. Similarly, *Per02* and *Per06* are both lawyers that are between 20 and 30 years old, and they both have flu.

```
<http://example.com/resource/Person/Per01> a foaf:Person;
    foaf:age "2x"; ex:hasJob ex:Teacher; ex:hasHIV "true".
<http://example.com/resource/Person/Per02> a foaf:Person;
    foaf:age "2x"; ex:hasJob ex:Lawyer; ex:hasFlu "true".
<http://example.com/resource/Person/Per06> a foaf:Person;
    foaf:age "2x"; ex:hasJob ex:Lawyer; ex:hasFlu "true".
<http://example.com/resource/Person/Per07> a foaf:Person;
    foaf:age "2x"; ex:hasJob ex:Teacher; ex:hasHIV "true".
```

Listing 6: Excerpt of a generalisation of the original graph – $G_{[0,1]}$.

Among the three 2-RDFanonymous graphs, $G_{[0,1]}$ generalises *Age* by one level, $G_{[1,1]}$ generalises both *Job* and *Age* by one level, and $G_{[2,1]}$ generalises *Job* by two levels and *Age* by one level. Furthermore, since $G_{[1,1]}$ is generalisation of $G_{[0,1]}$, and $G_{[2,1]}$ is a generalisation of both $G_{[1,1]}$ and $G_{[0,1]}$, $G_{[0,1]}$ is the *k-RDFminimal* generalisation of G.

Since there are three generalisation graphs that are 2-RDFanonymous, by calculating *RDFprec* for each graph it can be determined which of the three graphs has minimal distortion. In this use case, RDFprec($G_{[0,1]}$) = 0.75, RDFprec($G_{[1,1]}$) = 0.58, and RDFprec($G_{[2,1]}$) = 0.42. Therefore, the *RDFminimal* distortion of a graph G that satisfies 2-RDFanonymity is $G_{[0,1]}$.

V. Conclusions and Future Work

This paper has presented a framework for the anonymisation of RDF data that addresses the particularities of the RDF specification and can help in preserving the privacy of entities of interest in RDF data sets. Such framework describes an anonymisation model, several anonymisation operations, and an anonymisation metric.

The anonymisation model presented in this paper helps in preserving the privacy of RDF data sets. However, in order to ensure the maximum possible protection of privacy in RDF data by implementing such anonymisation model, it is first necessary to correctly identify resources that include information about QID and sensitive attributes, which is a difficult task. In those cases when these resources are not correctly identified, anonymisation can lead to an overprotection of the RDF data (i.e., lowering the precision of the anonymised data) or to a failure in ensuring the privacy of the RDF data.

The anonymisation model presented in this paper, although it ensures the protection of privacy to a certain level, is vulnerable to different kinds of attacks, such as in those cases when the order of entities of interest can compromise their privacy, or when subsequent releases of the same private information take place. Therefore, future work includes additional formalisations and recommendations on the implementation of the model to address those potential issues. These formalisations can be based on future case studies that will investigate the strength of the anonymisation model.

The framework for the anonymisation of RDF data described in this paper presents an initial effort towards the privacy protection of RDF data. Therefore, one line of future work consists of enriching the anonymisation framework described in this paper with other anonymisation models, besides the one presented in this paper, which can be based on the already existing set of models developed for tabular data. Furthermore, future work consists of including into the RDF anonymisation framework additional anonymisation metrics, and development of anonymisation algorithms. A variety of anonymisation models, metrics, and algorithms can help in achieving better privacy protection and will provide a comprehensive anonymisation framework for the privacy protection of RDF data.

Acknowledgments

This work is supported by the 4V project (TIN2013-46238-C4-2-R), funded by the Spanish Ministry of Economy and Competitiveness, and by the FPU grant (FPU2012/04084) of the Spanish Ministry of Education, Culture and Sport.

References

[1] L. Sweeney, "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 557–570, 2002.

[2] L. Sweeney, "Uniqueness of simple demographics in the U.S." Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Tech. Rep., 2000.

[3] G. Klyne, J. J. Caroll, and B. McBride, "RDF 1.1 Concepts and Abstract Syntax. Available online: http://www.w3.org/TR/rdf11-concepts/. Last retrieved on 10.12.2014." World Wide Web Consortium, Tech. Rep., 2014.

[4] B. C.M. Fung, K. Wang, A. Wai-Chee Fu, and P. S. Yu, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 2010.

[5] P. Samaratiy and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," SRI International, Tech. Rep., March. 1998.

[6] M. Ashwin, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), Atlanta, GA, USA, pp. 24–35*, 2006.

[7] P. Samaratiy, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, pp. 1010–1027, 2001.

[8] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, pp. 571–588, 2002.

[9] A. Gkoulalas-Divanis, S. Kotoulas, L. Vanessa, and M. L. Sbodio, "Guaranteeing anonymity in Linked Data graphs," International Patent PCT/US2014/033 261, 2014.