RDF/XML SOURCE DECLARATION

Fabien Gandon, Virginie Bottolier, Olivier Corby, Priscille Durville
INRIA - Edelweiss
2004 rt des Lucioles, BP93, 06902, Sophia Antipolis, France
Fabien.Gandon@sophia.inria.fr

ABSTRACT

When querying or reasoning on metadata from the semantic web, the source of this metadata can be of great importance. While the SPARQL query language provides a keyword to match patterns against named graphs, the RDF data model focuses on expressing triples. In many cases it is interesting to augment these RDF triples with the notion of a source for each triple (or set of triples), typically an IRI specifying their real or virtual origin. This article proposes and discusses an RDF/XML syntax extension providing an attribute to specify the source of triples in an RDF/XML representation.

KEYWORDS

Semantic Web, RDF, SPARQL, Source.

1. INTRODUCTION

When querying or reasoning on metadata from the semantic web, the source of this metadata can be of great importance. The Resource Description Framework or RDF (Klyne et al, 2004) is a general-purpose language for representing data and metadata on the web and it has an XML syntax called RDF/XML (Beckett D. and McBride B., 2004). The formal grammar for the syntax is annotated with actions generating triples of the RDF graph. In SPARQL (Prud'hommeaux E. and Seaborne A., 2007) when querying a collection of graphs, the GRAPH keyword is used to match patterns against named graphs. However the RDF data model focuses on expressing triples with a subject, predicate and object and neither it nor its RDF/XML syntax provide a mechanism to specify the source of each triple. A typical means would be an XML syntax to associate to the triples encoded in RDF/XML an IRI specifying their origin. This article proposes an extension of the syntax (a single attribute) to specify for these triples represented in RDF/XML the source they should be attached to.

2. SOURCE DECLARATION ATTRIBUTE IN RDF/XML

In SPARQL (Prud'hommeaux E. and Seaborne A., 2007) when querying a collection of graphs, the GRAPH keyword is used to match patterns against named graphs. GRAPH can provide an IRI to select one graph or use a variable which will range over the IRI of all the named graphs in the query's RDF dataset. The query in Figure 1 matches two graph patterns against each of the named graphs in the dataset and form solutions which have the `?srcname` and `?srctitle` variables bound to IRIs of the graph being matched.

```
01. PREFIX dc: <http://purl.org/dc/elements/1.1/>
02. PREFIX foaf: <http://xmlns.com/foaf/0.1/>
03. SELECT ?srcname ?name ?srctitle ?title
04. WHERE
05. {
06.   GRAPH ?srctitle  {  ?doc dc:title ?title .
07.                      ?doc dc:creator ?author  }
08.   GRAPH ?srcname   {  ?author foaf:name ?name  }
09. }
```

Figure 1. A SPARQL query on a dataset

Unfortunately the syntax of a SPARQL source has no equivalent in terms of the RDF syntax. We propose here a mechanism to standardize the declaration of sources in an RDF graph serialized in RDF/XML.
Using the Corese SPARQL engine (Corby O. et al., 2006), we implemented and tested an extension of the RDF/XML syntax: an attribute `cos:graph` may be inserted in an RDF/XML document to specify a source IRI. The value of this attribute is interpreted as an IRI Reference. The source IRI of a triple is:

1. the source IRI specified by a `cos:graph` attribute on the XML element encoding this triple, if one exists, otherwise
2. the source IRI of the element's parent element (obtained following recursively the same rules), otherwise
3. the base IRI of the document.

The base IRI of a document entity or an external entity is determined by RFC 2396 rules, namely, that the base IRI is the IRI used to retrieve the document entity or external entity. In other words, if no source is specified, the URL of the RDF/XML document is used as a default source.

The scope of a source declaration extends from the beginning of the start-tag in which it appears to the end of the corresponding end-tag, excluding the scope of any inner source declarations. Such a source declaration applies to all elements and attributes within its scope. In the case of an empty tag, the scope is the tag itself. Only one source can be declared as attribute of a single element.

Thus the `cos:graph` attribute can be used on any node element or property element to indicate that the included content is from a given source IRI. The most specific in-scope source present (if any) is applied.

We allow explicitly null sources: the `cos:graph=""` form indicates the absence of a source identifier so the associated source will explicitly be null and the base IRI of the document won't even be considered.

The RDF/XML syntax extension proposed here turns triples into quadruples with the new forth term being the IRI of the source of the triple. For instance consider an RDF graph stating that a resource has a title ("RDF Semantics") and a creator and that this creator is of type Person and has a name ("Patrick Hayes") and a mailbox ("mailto:phayes@ihmc.us") ; Figure 2 shows such a graph augmented with two occurrences of the `cos:graph` attribute. It results in having all the triples about the person in the source `http://www.ihmc.us` including the type declaration as a `foaf:Person`.

```
01. <rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
02.    xmlns:foaf="http://xmlns.com/foaf/0.1/"
03.    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04.    xmlns:cos="http://www.inria.fr/acacia/corese#"
05.    cos:graph="http://www.w3.org">
06.   <rdf:Description rdf:about="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
07.    <dc:title>RDF Semantics</dc:title>
08.    <dc:creator>
09.      <foaf:Person rdf:about="http://www.ihmc.us/users/user.php?UserID=42"
10.        cos:graph="http://www.ihmc.us" >
11.       <foaf:name>Patrick Hayes</foaf:name>
12.       <foaf:mbox rdf:resource="mailto:phayes@ihmc.us"/>
13.     </foaf:Person>
14.    </dc:creator>
15.   </rdf:Description>
16. </rdf:RDF>
```

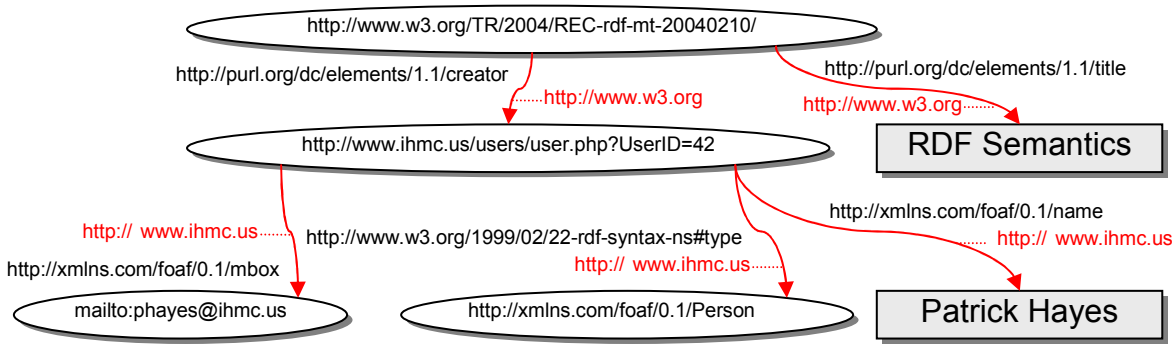

Figure 2. Example of the usage of the `cos:graph` attribute to specify two sources.

Quadruples resulting from the parsing of this file would be:

<http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> dc:title "RDF Semantics" **<- http://www.w3.org**

<http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> dc:creator <http://www.ihmc.us/users/user.php?UserID=42> **<- http://www.w3.org**

<http://www.ihmc.us/users/user.php?UserID=42> rdf:type foaf:Person **<- http://www.ihmc.us**

<http://www.ihmc.us/users/user.php?UserID=42> foaf:name "Patrick Hayes" **<- http://www.ihmc.us**

<http://www.ihmc.us/users/user.php?UserID=42> foaf:mbox <mailto:phayes@ihmc.us> **<- http://www.ihmc.us**

3. SOME PARTICULAR CASES

Generally speaking this sections shows that it is dangerous to change sources around blank nodes: following RDF and SPARQL semantics, a blank node must belong only to one source ; thus changing sources on properties of a blank node will result in splitting the blank node into several blank nodes, one for each source. To exemplify this point, this section walks you through the different particular cases of RDF graphs.

Blank nodes in descriptions with sources declaration can lead to surprising results. They should be used with care. In particular a blank node referenced in two (or more) different sources is interpreted as two (or more) different blank nodes, one for each source. The examples in Figures 3 and 4 lead to the creation of two blank nodes representing the person but with different attributions of the properties.

In cases where in one file we have two blank nodes with the same node ID and the same source there should be only one resulting blank node. Following RDF specifications, if the same node ID and the same source are used for a blank node in different files, it results in the creation of different blank nodes. Finally if the source is not effectively changed (e.g. several `cos:graph` are used on properties of a blank node but with the same IRI / source identifier) then only one blank node is created in the source.

Containers usually make use of explicit blank nodes. Changing the source on the blank node of a container or on one of its `rdf:li` properties will result in the creation of several blank nodes. This is true for `rdf:Bag`, `rdf:Alt` and `rdf:Seq`.

```
01. <rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
02.   xmlns:foaf="http://xmlns.com/foaf/0.1/"
03.   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04.   xmlns:cos="http://www.inria.fr/acacia/corese#"
05.   cos:graph="http://www.w3.org">
06.  <rdf:Description rdf:about="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
07.   <dc:title>RDF Semantics</dc:title>
08.   <dc:creator>
09.     <foaf:Person cos:graph="http://www.ihmc.us">
10.      <foaf:name>Patrick Hayes</foaf:name>
11.      <foaf:mbox rdf:resource="mailto:phayes@ihmc.us" />
12.    </foaf:Person>
13.   </dc:creator>
14.  </rdf:Description>
15. </rdf:RDF>
```

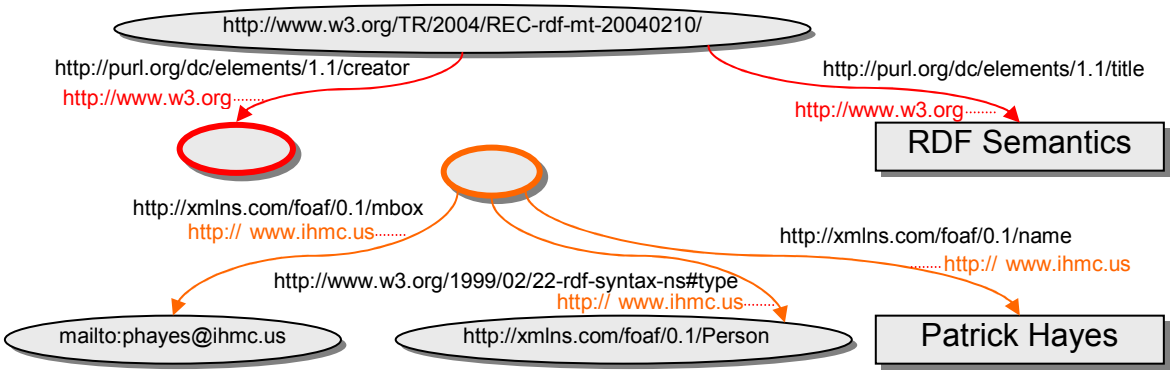

Figure 3. Blank node with a different source

```
01. <rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
02.   xmlns:foaf="http://xmlns.com/foaf/0.1/"
03.   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
04.   xmlns:cos="http://www.inria.fr/acacia/corese#"
05.   cos:graph="http://www.w3.org">
06.  <rdf:Description rdf:about="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
07.   <dc:title>RDF Semantics</dc:title>
08.   <dc:creator>
09.     <foaf:Person>
10.      <foaf:name>Patrick Hayes</foaf:name>
11.      <foaf:mbox rdf:resource="mailto:phayes@ihmc.us" cos:graph="http://www.ihmc.us"/>
12.    </foaf:Person>
13.   </dc:creator>
14.  </rdf:Description>
15. </rdf:RDF>
```

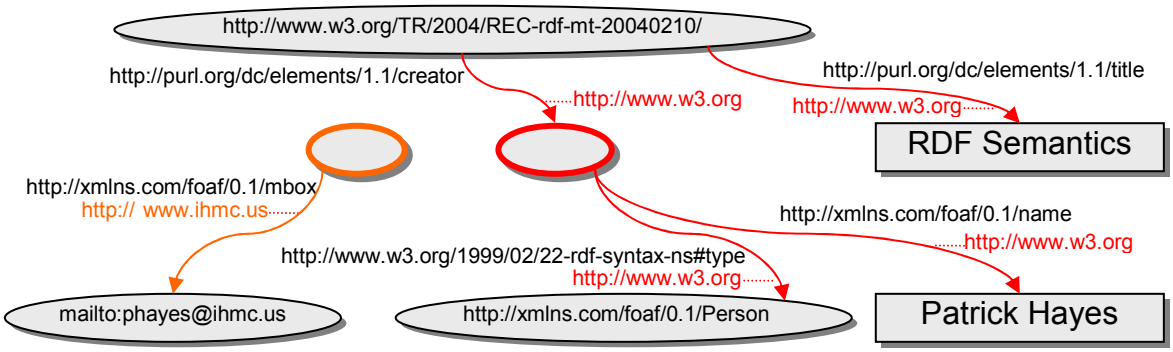

Figure 4. Blank node with a different source on one of its properties

Collections make use of implicit blank nodes. Changing the source on the description members of a collection does not cause any problem as long as these descriptions are ***not*** blank nodes themselves.

Structured values usually make use of implicit blank nodes generated by `rdf:parseType`. One should be careful in changing the source associated to a property of a value.

Triples generated by the reification of a triple belong to the same source as the original triple.

4. SOURCES AS RESOURCES

The IRI of a source can itself be the subject or object of RDF triple (Figure 5). Syntactically it is even possible to use the same IRI for a property and a source. However it is considered a bad practice and if a source is a described resource in OWL the OWL semantics would enforce that its IRI is not used to identify a property.

```
01. <rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/"
02.   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
03.   xmlns:cos="http://www.inria.fr/acacia/corese#"
04.   cos:graph="http://ns.inria.fr" >
05.   <foaf:Person rdf:about="http://ns.inria.fr/fabien.gandon">
06.     <foaf:name>Fabien Gandon</foaf:name>
07.     <foaf:mbox rdf:resource="mailto:Fabien.Gandon@sophia.inria.fr"/>
08.   </foaf:Person>
09.   <foaf:Organization rdf:about="http://ns.inria.fr" >
10.     <foaf:name>INRIA</foaf:name>
11.     <foaf:member rdf:resource="http://ns.inria.fr/fabien.gandon"/>
12.   </foaf:Organization>
13. </rdf:RDF>
```

Figure 5. Annotation about the IRI of a source

Sources can be used in RDFS schemas to specify the source of the definitions and in OWL schemas one can link the source of a schema to its namespace (Figure 6). Once again, one should be careful with blank nodes when annotating schemas, be them explicit or implicit. This is particularly true for some primitives in OWL such as restrictions, unions, intersections.

```
01. <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
02.          xmlns:cos="http://www.inria.fr/acacia/corese#"
03.          xmlns:owl="http://www.w3.org/2002/07/owl#"
04.          xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
05.          xmlns="http://www.w3.org/2000/01/rdf-schema#"
06.          cos:graph="http://ns.inria.fr/2006/05/research_onto.rdfs"
07.          xml:base="http://ns.inria.fr/2006/05/research_onto.rdfs">
08. <owl:Ontology rdf:about="http://ns.inria.fr/2006/05/research_onto.rdfs" >
09.    <rdfs:label>research ontology</rdfs:label>
10.    <rdfs:comment>defines primitives to describe research activities</rdfs:comment>
11.    <rdfs:comment cos:graph="http://www.w3.org">RDFS_VALID</rdfs:comment>
12.    <owl:versionInfo>1.3</owl:versionInfo>
13. </owl:Ontology>
14. </rdf:RDF>
```

Figure 6. The case of an OWL Ontology annotation

5. CONCLUSION

As far as we know, the closest works to what we presented here are related to the RDF contexts issue raised during the redesign of RDF in 2002[1]. A suggestion was made that the concept of context was missing from RDF but the RDFCore WG decided to postpone this issue to a future version of RDF[2]. This issue was raised by (Klyne, 2002) who considered scoping assertions by context but with the intent to express relations between these contexts such as negation, implication, etc. In addition the article suggested that the structure of contexts might be encoded in RDF using the RDF reification. In an other attempt, (De Vos, 2001) proposed to annotate property elements with the attribute `rdf:parseType="Statements"` to indicate that their content is considered the same as the content model of the `rdf:RDF` element but that the statements are in a separate context. However in this approach one could specify the naming (identificator) that context. In N3 (Berners-Lee, 2006) explains that RDF has no datatype allowing a graph to be a literal value. So he proposed that N3 extends this notation to allow a graph itself to be referred to as a formula i.e. a list of statements between braces whose meaning is the logical conjunction of the statements in the list. A formula may then be annotated with properties but there is no mechanism to associate a name or identificator to this formula. This specification was tested with a modified version of the SPARQL engine CORESE (Corby O. et al., 2006) and is available in version 2.4 of CORESE. This specification was driven by use cases from several of our projects. However there is one case we left out of the scope: the case where one wants to attach several sources to a triple. We did not find a good syntax for this case and we did not investigate what it would imply in terms of SPARQL querying.

REFERENCES

Beckett D. and McBride B., 2004. RDF/XML Syntax Specification (Revised), *W3C Recommendation*, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

Berners-Lee T., 2006, Notation 3, http://www.w3.org/DesignIssues/Notation3.html

Corby O. et al., 2006. Searching the Semantic Web: Approximate Query Processing based on Ontologies. In *IEEE Intelligent Systems*, January/February 2006, Vol. 21, No. 1, pp 20-27

De Vos A., 2001, RDF Difference Models Representing the Difference between two RDF Models, Langdale Consultants

Klyne et al, 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax., *W3C Recommendation*, 10 February 2004, http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/.

Klyne G., 2002, Circumstance, provenance and partial knowledge Limiting the scope of RDF assertions, http://www.ninebynine.org/RDFNotes/UsingContextsWithRDF.html

Prud'hommeaux E. and Seaborne A., 2007. SPARQL Query Language for RDF. *W3C Candidate Recommendation*, 14 June 2007, http://www.w3.org/TR/2007/CR-rdf-sparql-query-20070614/.

[1] http://www.w3.org/2000/03/rdf-tracking/#rdfms-contexts

[2] http://lists.w3.org/Archives/Public/www-rdf-comments/2002JulSep/0096.html