Name That Graph or the need to provide a model and syntax extension to specify the provenance of RDF graphs

Fabien Gandon, Olivier Corby

Name That Graph

or the need to provide a model and syntax extension to specify the provenance of RDF graphs.

Fabien Gandon and Olivier Corby, INRIA, W3C Member

ABSTRACT

When querying or reasoning on metadata from the semantic web, the source of this metadata as well as a number of other characteristics (date, trust, etc.) 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 provenance for each triple (or set of triples), typically an IRI specifying their real (e.g. author) or virtual origin (e.g. inference engine and entailment regime). This position paper extends on a member submission from INRIA (RDF/XML Source Declaration [1]) expressing the importance for us to provide a standard mechanism in RDF to name graphs. It proposes and discusses an RDF/XML syntax extension providing an attribute to specify the provenance of triples in an RDF/XML representation.

INTRODUCTION

When querying or reasoning on metadata from the semantic web, the provenance of this metadata can be of great importance. The Resource Description Framework or RDF [11] is a general-purpose language for representing data and metadata on the web and it has an XML syntax called RDF/XML [12]. The formal grammar for the syntax is annotated with actions generating triples of the RDF graph. In SPARQL **Erreur ! Source du renvoi introuvable.** 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 the model nor its 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 triples represented in RDF/XML the IRI of the source they should be attached to as a provenance and tracing mechanism.

THE MISSING LINK

In SPARQL **Erreur ! Source du renvoi introuvable.** 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 IRIs of all the named graphs in the triple store. The query in Figure 1 matches two graph patterns against each of the named graphs in the triple store, and forms 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.

RDF provides constructs to write reification quads [14] but in RDF asserting the reification is not the same as asserting the original statement, and neither implies the other. Moreover reification expands the initial triple into a total of five triples (a triple plus a reification quad) and the link between the initial triple and its reification quad is not maintained.

The attribute `rdf:ID` can also be used in a property element to produce a reification of the triple that the property element generates and assert it at the same time. However this mechanism remains at the level of triples and there is nothing in the resulting triples that explicitly identifies the original triple and links it to the reification quad. RDF provides no way to associate the subject of the reification triples with an individual triple.

Likewise, statements can be made using the URI of a document as commonly done by annotations in OWL. In an ad-hoc application-dependent understanding, those statements could be interpreted as if they were to be distributed over all the statements in the document. But here again we are outside RDF and relying on likening the document to its asserted content does not sound like a good practice.

Therefore, nowadays, associating specific URIrefs with specific statements has to be done using mechanisms outside RDF and is one of the motivations behind RDF 2.0.

NAMED GRAPHS, NESTED GRAPHS AND CONTEXT

Carroll et al. [15] remarked that RDF does not provide any operational means, apart for the aspects mentioned above, for making statements about graphs and relations between graphs. As a solution to this problem they proposed Named Graphs in RDF to allow publishers to communicate the assertional intent their assertions and sign them. The fact that it is often useful to embody social acts with some record clearly resonates with all the scenarios where questions are raised about the assertions manipulated by the system (provenance, certification, dating, quality, range, target, rights, etc.).

Several authors before proposed to transform RDF triples into quads [16][17][18][19] appending to them an additional URI Reference, blank node or ID. The definition of Carroll et al. [15] is deliberately simpler than [20] and [21] and states that a “Named Graph is an RDF graph which is assigned a name in the form of a URIref. The name of a graph may occur either in the graph itself, in other graphs, or not at all. Graphs may share URIrefs but not blank nodes." [15].

In addition to these contributions, related works can also be found in the many other graph-based knowledge representation formalisms, Conceptual Graphs being one of them as an historical descendant of semantic networks. We believe it will be important to survey previous contributions such as the notion of context (Chapter 5 in [2], example 5 in [3], section 4 in [4]) or the notion of nested graphs [5][6][7][8][9].

SOURCE DECLARATION ATTRIBUTE IN RDF/XML

To serialize named graphs, Carroll et al. used TriX and TriG [15] but noted that RDF/XML is the deployed base. Therefore, we proposed in the W3C Member Submission “RDF/XML Source Declaration” [1] an XML syntax to associate to the triples encoded in RDF/XML an IRI specifying their origin. This extension uses a single attribute to specify for these triples represented in RDF/XML the IRI naming the source graph they should be attached to.

Using the Corese SPARQL engine [22], we implemented and tested this extension of the RDF/XML syntax: an attribute `cos:graph` may be inserted in an RDF/XML document to name a graph. 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 [23], 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. A limitation we will come back to is that 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 graph containing the triple. Let us consider the following RDF graph stating that a resource has a title ("`RDF Source`") and a creator and that this creator is of type Person and has a name ("`Fabien Gandon`") and a mailbox ("`mailto:fgandon@inria.fr`"). 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 graph named `http://www.inria.fr` 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-sop.inria.fr/edelweiss/fabien/docs/w3c/rdfsource/rdfsource.html">
07.   <dc:title>RDF Source</dc:title>
08.   <dc:creator>
09.     <foaf:Person rdf:about="http://ns.inria.fr/fabien.gandon/foaf#me"
10.       cos:graph="http://www.inria.fr" >
11.      <foaf:name>Fabien Gandon</foaf:name>
12.      <foaf:mbox rdf:resource="mailto:fgandon@inria.fr"/>
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-sop.inria.fr/edelweiss/fabien/docs/w3c/rdfsource/rdfsource.html> dc:title "RDF Source" **<http://www.w3.org>**
<http://www-sop.inria.fr/edelweiss/fabien/docs/w3c/rdfsource/rdfsource.html> dc:creator <http://ns.inria.fr/fabien.gandon/foaf#me> **<http://www.w3.org>**
<http://ns.inria.fr/fabien.gandon/foaf#me> rdf:type foaf:Person **<http://www.inria.fr>**
<http://ns.inria.fr/fabien.gandon/foaf#me> foaf:name "Fabien Gandon" **<http://www.inria.fr>**
<http://ns.inria.fr/fabien.gandon/foaf#me> foaf:mbox <mailto:fgandon@inria.fr> **<http://www.inria.fr>**

SOME EXAMPLES OF SUCH AN EXTENSION IN USE

This specification and its implementation were driven by use cases from several of our projects and tested in CORESE, a SPARQL and RDFS/OWL Light engine.

Named graphs to embody the social act of tagging

The need expressed by Carroll et al. [15] to embody social acts with some record naturally applies to the case of representing social tagging. In the NiceTag model and experiment [24], to model tag actions we defined a subclass of named graphs (modeled as `rdfg:Graph` [15]) called `TagAction` which embodies one single act of tagging (see fig. 3 below). The triples contained in the named graph represent the link, modeled with the property `isRelatedTo`, between an instance of the class `irw:Resource` and a sign (modeled as an instance of `rdfs:Resource`).

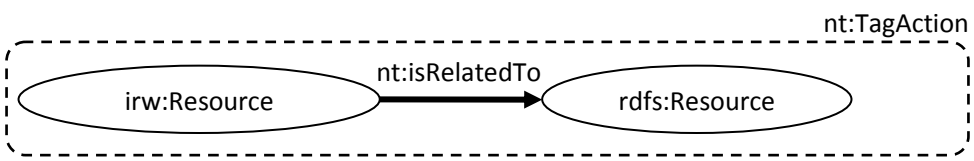

Fig. 3. A Tag Action as a named graph

The IRI of the named graph of the tagging act identifies a resource that can be described and typed. To account for the nature of the different possible tag actions, we defined subclasses of the `TagAction` class. For instance we type the named graph to distinguish between tagging performed by machines (`AutoTagAction`), from tagging performed by humans (`ManualTagAction`), from more complex types of tagging as those involving machine tags (`MachineTagAction`). In addition we can attach properties to describe the place where tag actions are stored, the account of the user who tagged, the date the tagging act occurred, etc.

Design Pattern for Contextual Metadata

In several e-Science projects we used named graphs to represent and query contextual metadata [25][22]. For instance, evidence-based reasoning requires being able to differentiate assertions considered as universally true and assertions which are concurrent hypothesis or interpretations. Thus we use named graphs when annotating experiments (e.g. in biology, Sealife Project) or analysis (e.g. in geology, e-Wok project). Named graphs are used to represent different contexts within which alternative metadata can be described.

Naming the graphs also allowed us to hierarchically organize the RDF datasets, based on RDFS entailment. When considering RDF datasets as contexts, the root of the hierarchy contains the triples that are true in any context below it *i.e.* any other node of the hierarchy entails it. The other nodes of the hierarchy represent specific contexts; each one recursively inherits and specializes the triples of its ancestors. Each node then provides a different context for querying and reasoning. When a hypothesis is tested (as a SPAQL query), the context of the test is specified by the name of the graph to be used.

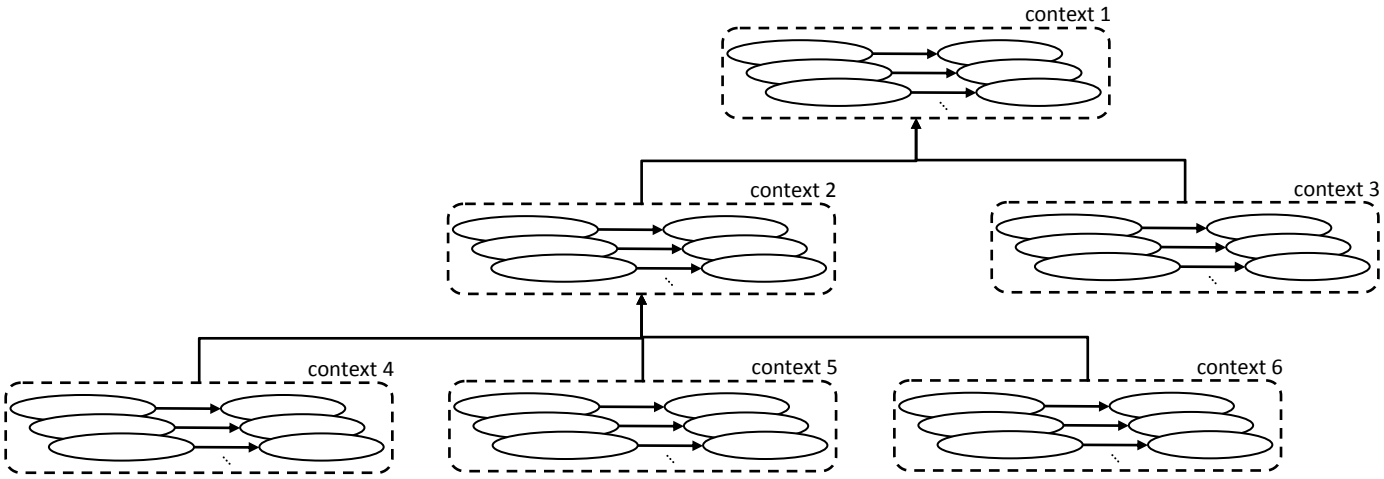

Fig. 4. Hierarchically organized named graphs providing nested contexts for querying and reasoning

Other use cases under consideration

We currently consider the ability to name graphs in order to:

- trace the inferences that enriched a triple store to undo some reasoning when the base is updated.
- date assertions and introduce some temporal aspects in querying and reasoning over our triple store.
- assign a definition (a graph) to an IRI (its name).
- capture context in representing the result of natural language processing and knowledge extraction.
- accommodate and manage different points of view.
- assist distributed storage and query routing.

DISCUSSION

Beyond the simple proposal of this position paper, a number of particular cases are addressed in details in the member submission [1]. In particular we underlined that it is dangerous to change sources around blank nodes: following RDF and SPARQL semantics, and the named graph model [15], each blank node can belong to only one named graph; thus changing sources on properties of a blank node results in splitting the blank node into several

blank nodes, one for each source. To exemplify this point, a complete section of the member submission [1] walks the reader through a collection of special cases.

This proposal also left out the important case where one wants to attach several source graphs to a triple. We did not find a good syntax for this case and from the model stand point may be the notion of surface introduced by Pat Hayes in his invited talk at ISWC 2009 would be more adequate.

From the deployment stand point such extension introduces problems of backward compatibility. The generated triples are the same whether the parser understands source attributes or not. However the graph names will be different and this may lead to interoperability issues for instance for SPARQL Query using `GRAPH` or `FROM NAMED` clauses.

Finally, work on named graphs should most probably coordinate with the work of the Incubator Group on Provenance [10].

To conclude, in our opinion, the issue of Named Graphs, their semantics and syntax, as well as possible new XML syntaxes for RDF are seminal work items for an RDF 2.0 Working Group.

REFERENCES

[1] Gandon F., Bottolier V., Corby O., Durville P.: RDF/XML source declaration, W3C Member Submission. http://www.w3.org/Submission/rdfsource/ (2007)

[2] John F. Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. Actual publication date, 16 August 1999, ISBN 0-534-94965-7

[3] Conceptual Graph Examples, John F. Sowa, http://www.jfsowa.com/cg/cgexampw.htm#Ex_5

[4] Conceptual Graphs, John F. Sowa, Chapter 5 of the Handbook of Knowledge Representation, ed. by F. van Harmelen, V. Lifschitz, and B. Porter, Elsevier, 2008, pp. 213-237

[5] D. Genest and E. Salvat. A Platform Allowing Typed Nested Graphs: How CoGITo Became CoGITaNT. In Proceedings of the 6th International Conference on Conceptual Structures, Lecture Notes in AI. Springer, 1998.

[6] Two FOL Semantics for Simple and Nested Conceptual Graphs, G. Simonet, Conceptual Structures: Theory, Tools and Applications, LNCS 1453, 1998

[7] Michel Chein and Marie-Laure Mugnier. Positive nested conceptual graphs. In Proceedings of ICCS '97, volume 1257 of LNAI, pages 95-109, Springer, 1997.

[8] M. Chein, M.L. Mugnier, and G. Simonet. Nested graphs: A graph-based knowledge representation model with fol semantics. In Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), Trento, Italy, June 1998.

[9] A Preller, M.L. Mugnier, and M. Chein. Logic for Nested Graphs. Computational Intelligence Journal, (CI 95-02-558), 1996.

[10] W3C Provenance Incubator Group, http://www.w3.org/2005/Incubator/prov/

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

[12] 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/

[13] Prud'hommeaux E. and Seaborne A., 2008. SPARQL Query Language for RDF W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/

[14] RDF Reification, RDF Primer W3C Recommendation 10 February 2004, Frank Manola, Eric Miller http://www.w3.org/TR/rdf-primer/#reification

[15] Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In 14th Int. Conference on World Wide Web WWW, pages 613-622, New York, NY, USA, 2005. ACM

[16] D. Beckett. Redland notes - contexts. http://www.redland.opensource.ac.uk-/notes/contexts.html 2003

[17] E. Dumbill. Tracking provenance of rdf data. Technical report, ISO/IEC, 2003

[18] Intellidimension. Rdf gateway - database fundamentals. http://www.intellidimension.com/pages/dfgateway/dev-guide/db/db.rsp, 2003.

[19] R. MacGregor and I.-Y. Ko. Representing contextualized data using semantic web tools. In Practical and Scalable Semantic Systems (ISWC workshop), 2003

[20] R. M. R. Guha and R. Fikes. Contexts for the semantic web. In Int. Sem Web Conf, ISWC, 2004.

[21] M. Sintek and S. Decker. Triple - a query, inference, and transformation language for the semantic web. In Intl. Sem. Web Conf., ISWC, 2002.

[22] Olivier Corby, Web, Graphs and Semantics. ICCS 2008: 43-61

[23] Uniform Resource Identifiers (URI): Generic Syntax, T. Berners-Lee, R. Fielding, U.C. Irvine, L. Masinter, August 1998, http://www.ietf.org/rfc/rfc2396.txt

[24] Limpens, F, Monnin A, Laniado D, Gandon F., NiceTag Ontology: tags as named graphs, International Workshop in Social Networks Interoperability, 2009. See also http://ns.inria.fr/nicetag/2009/09/25/voc.html

[25] RDF/SPARQL Design Pattern for Contextual Metadata Olivier Corby, Catherine Faron-Zucker, IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, USA, November 2007