Named Graphs as a Mechanism for Reasoning about Provenance

E. Rowland Watkins[1] and Denis A. Nicole[2]

[1] IT Innovation Centre, 2 Venture Road, Chilworth Science Park, Southampton, UK, SO16 7NP, erw@it-innovation.soton.ac.uk,
[2] School of Electronics & Computer Science, University of Southampton, Southampton, UK, SO17 1BJ, dan@ecs.soton.ac.uk

Abstract. Named Graphs is a simple, compatible extension to the RDF abstract syntax that enables statements to be made about RDF graphs. This approach is in contrast to earlier attempts such as RDF reification, or knowledge-base specific extensions including quads and contexts. In this paper we demonstrate the use of Named Graphs and our experiences developing new kinds of semantic web application that build on Named Graphs for digital signatures, provenance, and semantic reasoning. We present a working example based on the Named Graphs for Jena (NG4J) API, from which we developed a semantic version control system for Software Engineering capable of reasoning about Named Graph-based provenance. We go on to discuss the implications of Named Graphs for Description Logics and semantic inference strategies.

1 Introduction

The Semantic Web is intended to move the current "textual" Web into a Web of Knowledge. RDF [1] provides a way to describe relations between Web resources as a graph; it records these relationships as (subject, verb, object) triples. Description Logic languages based on standards such as RDFS and OWL, model the interrelations between types of resource. RDF defines how we may merge a set of graphs into one, but does not provide mechanisms for showing relationships between graphs. This ability, for example, to attribute provenance to RDF will become important as the Semantic Web grows.

RDF defines the term reification in its formal semantics [2] as a means of recording descriptions of triples. Unfortunately, reified statements cannot be used in semantic inferences, and are not asserted as part of the underlying knowledge-base. The result of this is that Semantic Web practitioners must look elsewhere for provenance recording mechanisms.

Defining provenance and recording it is a non-trivial problem since the level of recording and its intended purpose varies from one application to another. Despite this, several solutions have been proposed to address the provenance problem in RDF, most of which are compatible with, and extend, RDF abstract syntax. These include Contexts, Quads and Named Graphs.

Named Graphs, introduced by Carroll et al. [3], defines an extension that labels RDF graphs; in this paper we describe how we have used Named Graphs to reason about provenance using existing Description Logics.

The remainder of this paper provides an overview of Named Graphs and our experiences of using them. Section 2 describes earlier work and how Named Graphs can naturally record provenance. Section 3 describes an example online collaborative tool that uses Named Graphs extensively as a framework for version control. Section 4 recounts our experience of Named Graphs and the challenges faced in their deployment. We discuss related work in 5.

2 Recording Provenance

Whilst the earlier work is of benefit for the development of domain-specific provenance frameworks, it does little to show how provenance can be managed uniformly.

2.1 Provenance in Bioinformatics

The ^{my}Grid project[3] has developed Grid middleware to meet the needs of bioinformatics. In this domain, it was essential to be able to capture and manipulate provenance information [4, 5]. The project takes provenance records from sources such as the Freefluo[4] workflow orchestration tool and uses an ontology to annotate these provenance records for future analysis.

2.2 RDF Reification

RDF reification was intended as a mechanism for making provenance statements and other statements about RDF triples. Each triple is described with a special vocabulary which includes `rdf:Statement`.

Several problems become apparent when we attempt to assert provenance in this way. The key issue is that the presence of a reified triple in the knowledgebase is unrelated to the presence of the triple itself; thus including the reification does not of itself assert the triple. If we choose to assert each triple as well as its reification, then it is asserted unconditionally and this triple is not bound to the reification.

One major consequence of these problems is that it is not possible to reason about triples in the context of their provenance through the RDF reification mechanism. If the reified triple is not bound to an asserted (or not as the case may be) triple, then RDF reification is of no real use in the recording of provenance.

[3] http://www.mygrid.org.uk/.
[4] http://freefluo.sourceforge.net/.

```
:G1 {
   :Bob foaf:mbox <mailto:bob@example.org>.
   :G2 foaf:maker "Rowland Watkins".
}
:G2 {
   :Bob foaf:mbox <mailto:bob2@example.com>.
   :G2 dcterms:created "25-7-2005".
}
```

Fig. 1. Self-Referencing and Cross-Referencing Named Graphs.

2.3 Named Graphs

Named Graphs provide a natural way to record provenance. Each graph is potentially labelled by a URI, which can then be referenced by other Named Graphs. Fig. 1 depicts two Named Graphs using the TriG syntax [6], where the first graph states that the second graph was made by Rowland Watkins, while the second graph self-references, stating its creation date.

Provenance with Named Graphs is not limited to simple assertions. Such assertions, whilst true according to the open world assumption, do not uniquely bind an owner to an assertion, or set of assertions. Cryptographic methods such as digital signatures offer one way to uniquely bind a security principal to a digital document and coupled with digital certificates add non-repudiation to signatures. Such methods are also a first step towards a basic level of trust on the Semantic Web.

RDF signatures have been realized as part of the Semantic Web Publishing framework (SWP), an extension to NG4J [7]. SWP defines an ontology that follows the XML Signature Recommendation. We create a canonical Named Graph [8] then hash it with an appropriate secure digest (SHA-1 in our case). This digest is placed in a special Named Graph called a Warrant Graph [3].

A Warrant Graph can contain any number of graph digests. Each digested graph is explicitly asserted by a known principal who possesses a digital certificate (X.509) or PGP key. The Warrant Graph asserts itself and signs itself with the principals credentials, certifying that not only did the principal make the assertion, but that the assertion has not been altered.

3 Application

As an example of how Named Graphs can be used as a mechanism for reasoning about provenance, we developed our document provenance ontology that describe resources in a software version control repository; this ontology became the basis for out online collaborative tool [9]. This was conceived as an alternative to systems such as CVS. Fig. 2 shows our ontology which makes maximum use

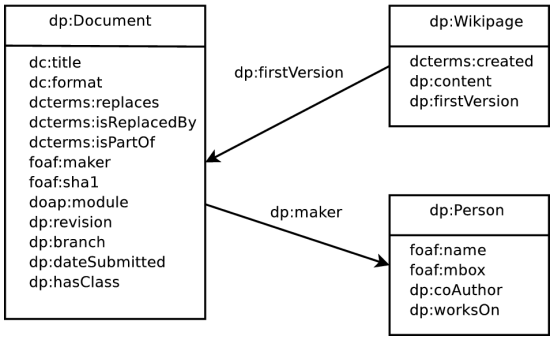

Fig. 2. Document Provenance Ontology.

of several other well known ontologies (DCMI[5], FOAF, DOAP) to help maintain interoperability.

Although our ontology appears small and simplistic it does capture all the information that we might expect in version control metadata, e.g. relationships to other versions (`dcterms:replaces`, `dcterms:isReplacedBy`), authorship (`foaf:maker`), and commit date (`dp:dateSubmitted`). By keeping our ontology small we reduce complexity and increase maintainability.

We have used our ontology as the basis for a semantic version control system; class instances serve a similar role to a relational database that we query and display in a WikiWikiWeb interface. While RDF is an interesting method for storing data, it is invariably slower than a relational database. An OWL-DL ontology, however, has a distinct advantage over a relational database since we are able to perform semantic inferences over our instance data; inferred data might tell us new information based on questions (queries) made by a developer to the version control system.

4 Discussion

Applying Named Graphs to our OWL-DL ontology allowed us to effectively partition metadata which could then be signed using our work on SWP. Each top-level class instance in the Document Provenance ontology is contained in a Named Graph, signed with a digital signature in a Warrant graph. The combination of digital signatures and Description Logic means at its base our online collaborative tool has two levels of internal verification: cryptographically verify the integrity of the our metadata; check the semantic consistency of our metadata using the OWL-DL Class and Property axioms.

In addition to OWL-DL consistency checking, we have used the Jena[6] inferencing engine because we are able to tailor its inference rules to suit our needs.

[5] The Dublin Core vocabularies used in our work are OWL-DL versions: http://protege.stanford.edu/plugins/owl/dc/.

[6] http://jena.sourceforge.net/inference/.

Custom rules have been used to answer complex queries in our online collaborative tool. Inferences over large semantic datastores is inevitably a slow process; our strategy has been rather than update inferences incrementally, they are run as required since they are rederivable.

5 Related Work

TRIPLE [10] adopts a Named Graph approach; however, it incorporates data representation and Horn-clause logic in the same syntax. It is intended as a rule language supporting applications that require RDF reasoning and transformation under different semantics. Its use of Horn-clause logic means it can be enacted by Prolog systems.

3Store [11], uses quads to track the provenance of triples which has been used in several novel applications including http://hyphen.info/. 3Store also supports RDFS entailment, although there does not appear to be any general purpose inference engine to date.

Reggiori et al. [12] use contexts as a means to record provenance in their RDFStore. They see contexts as an additional and orthogonal dimension to the RDF triple where each RDF statement is flagged as belonging to a specific context.

The Provenance Aware Service Oriented Architecture (PASOA) Project continues some of the work done by the ^{my}Grid project on data provenance. Its aim is to investigate the nature of provenance and reason about the accuracy of data and service in the e-Science domain. It has so far developed a provenance recording service, called PReServ [13], an implementation of the Provenance Recording Protocol (PReP) [14] developed by the PASOA project[7]. PASOA can and will serve as a valuable frontend for gathering provenance data into our reasoning mechanism.

6 Conclusion

Named Graphs provide a natural way to record provenance. They offer an alternative to RDF reification that is powerful and has been used in practice to associate digital signatures with graphs and reason about them.

Our work on semantic version control brings together the Semantic Web, digital signatures and the WikiWikiWeb, demonstrating Named Graphs to be of practical use. Not only does our solution have a query mechanism for displaying RDF in the Wiki, it also has an advanced semantic inference facility so that software developers can learn more about the software engineering process. Experience from this work has taught us a great deal about the effects Named Graphs have on Description Logic languages and semantic inference strategies.

Future work will see a grid service interface compatible with those provided by the Open Middleware Infrastructure Institute (OMII). Maven[8] integration

[7] http://www.pasoa.org/.
[8] http://maven.apache.org/.

would also be of benefit, integrating software project management with software version control.

References

1. G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax, 2004.
2. P. Hayes. RDF Semantics, 2004.
3. Jeremy J. Carroll, Christian Bizer, Pat Hayes and Patrick Stickler. Named Graphs, Provenance and Trust. In Proceedings of the 14th International World Wide Web Conference, Chiba, Japan, May 10-14, 2005.
4. Jun Zhao, Carole Goble, Mark Greenwood, Chris Wroe and Robert Stevens. Annotating, linking and browsing provenance logs for e-Science. In the Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Florida, USA, October 2003.
5. Jun Zhao, Carole Goble, Robert Stevens and Sean Bechhofer. Semantically Linking and Browsing Provenance Logs for e-Science. In International Conference on Semantics of a Networked World, Paris, France, P158-176, Lecture Notes in Computer Science, 2004. Springer.
6. Chris Bizer. The TriG Syntax, 2004.
7. Chris Bizer, Richard Cyganiak and Rowland Watkins. Named Graphs for Jena (NG4J) API. In The Second European Semantic Web Conference, Heraklion, Greece, 29 May - 1 June, 2005.
8. Jeremy J. Carroll. Signing RDF Graphs. In 2nd ISWC, volume 2870 of LNCS. Springer, 2003.
9. E. Rowland Watkins and Denis A. Nicole. Version Control in Online Software Repositories. In Proceedings of the 2005 International MultiConference in Computer Science & Computer Engineering Las Vegas, Nevada, USA June 27-30, 2005.
10. Michael Sintek, Stefan Decker, TRIPLE - A Query, Inference, and Transformation Language for the Semantic Web. In Proceedings of the First International Semantic Web Conference on The Semantic Web, p.364-378, June 09-12, 2002.
11. Stephen Harris and Nicholas Gibbins. 3Store: Efficient Bulk RDF Storage. In Proceedings of the First International Workshop on Practical and Scalable Semantic Systems, Sanibel Island, Florida, USA October 20, 2003.
12. Alberto Reggiori, Dirk-Willem van Gulik and Zavisa Bjelogrlic. Indexing and retrieving Semantic Web resources: the RDFStore model. In SWAD-Europe Workshop on Semantic Web Storage and Retrieval 13-14 November 2003, Vrije Universiteit, Amsterdam, Netherlands.
13. Paul Groth, Simon Miles, Weijian Fang, Sylvia C. Wong, Klaus-Peter Zauner, and Luc Moreau. Recording and Using Provenance in a Protein Compressibility Experiment. In Proceedings of the 14th IEEE International Symposium on High Performance Distributed Computing (HPDC'05), July 2005.
14. Paul Groth, Michael Luck, and Luc Moreau. A protocol for recording provenance in service-oriented Grids. In Proceedings of the 8th International Conference on Principles of Distributed Systems (OPODIS'04), Grenoble, France, December 2004.