Ontology Based Tracking and Propagation of Provenance Metadata

Miroslav Vacura and Vojtěch Svátek

Faculty of Informatics and Statistics,
University of Economics
W. Churchill Sq.4, 130 67 Prague 3,
Czech Republic
vacuram|svatek@vse.cz

Abstract. Tracking the provenance of application data is of key importance in the network environment due to the abundance of heterogeneous and controllable resources. We focus on ontologies as a mean of knowledge representation and present a novel approach to representation of provenance metadata in knowledge bases, relying on an OWL 2 design pattern. We also outline an abstract method of propagation of provenance metadata during the reasoning process.

1 Introduction

One of important features of any complex network environment is the multiplicity of information sources. This situation completely changes the information processing paradigm: in a conventional information system data come from a limited and usually relatively small number of sources. Sources of data are controllable and uncertainty regarding data reliability can be limited. Huge networks like World Wide Web, on other hand, consists of an enormous number of different information sources, which are usually completely uncontrollable and their reliability is usually questionable. If we use WWW data for the purposes of entertainment, this character of WWW data is not a problem, but if we intend to use the WWW for serious business or scientific applications, keeping track of the origins of data becomes necessary. Namely, when working with typical network applications, which usually process data from multiple WWW sources, data provenance is of key importance. However, until now, most web applications lack any data provenance features.

Buneman et al. [4, 5] defines *data provenance* as "the process of tracking and recording the origins of data and its movement between databases".

In following text we will focus on Semantic web [3] applications and context. However information about the origin of a piece of data and the process by which it arrived to information system is not only important in the case of Semantic Web applications. For many other types of applications this information is of critical importance too, like in the case of Molecular Biology or in the cases where legal or ethic issues are associated with the data involved [4].

In the context of the Semantic web, provenance information can be attached to RDF triples using ad hoc reification. Such solutions however make the reuse of such information hard, and also do not fit well to annotation of ontologies themselves. We therefore propose a solution based on a design pattern that uniformly captures provenance information for an ontology as well as for the data (RDF knowledge base) that are based on it.

The rest of the paper is organized as follows. Section 2 briefly discusses the state of the art in provenance (and similar metadata) representation in the Semantic web and presents the ontology pattern for representing provenance metadata, relying on the recent OWL 2 specification. Section 3 then proposes a mechanism for propagation of provenance metadata in ontologies. Finally, section 4 summarises the content.

2 Representation of Provenance Data in Ontologies

2.1 State of the Art

There are generally two ways of conceptualising provenance data in ontologies. One way is to include provenance information in the same ontology along with other information. The other way is to use a separate ontology for regular data and a separate ontology for provenance data. This distinction not only holds for representing provenance metadata but also for describing any kind of metadata, including metadata regarding certainty or relevance of information. The first approach is used for example in the COMM multimedia ontology [1], a comprehensive framework based on the DOLCE foundational ontology and the MPEG-7 standard, which is focused on describing multimedia data. Provenance data can be included in this ontology along with other descriptive information regarding given multimedia data, and its representation is based on the "Descriptions & Situations" design pattern. The second approach—using a separate ontology for describing metadata—is used for example for describing *relevance* of information [6] or in our recent work regarding representation of *uncertainty* [11].

Both these approaches can be used, based on context and situation. The first approach is, from our point of view, more appropriate when the whole ontology contains information that has the character of metadata. This is case of COMM multimedia ontology, where multimedia data are stored in individual files in the file system, while the ontology contains metadata describing this multimedia with different characteristics. Then provenance data is just one kind of metadata associated with multimedia data.

On the other hand, there are also situations when the ontology represents actual data and therefore it can be reasonable to use an additional metadata ontology to represent metadata information like the certainty of data, relevance of data or in our case provenance data. We don't see this distinction as obligatory, as it is always a design decision of developers of the ontology how they intend to represent these kinds of information in their ontology, and this decision de-

pends on concrete circumstances, domain and context of ontology that is being developed.

In this paper we present an instance of the second approach, assuming it is adequate in many cases.

2.2 OWL DL Setting

Formal basis for ontologies is provided by Description Logics (DL) [2]. In DL we understand an ontology as a triple $O =< K_R, K_T, K_A >$, where K_R is the role box (RBox), K_T is the terminology box (TBox), and K_A is the assertional box (ABox) – see [2] for detailed description of DL.

In reality an ontology O can be result of merging several other ontologies with different or same level of generality. Such ontology may be for example developed on the basis of some foundational ontology O^F, with two other domain ontologies O^{D_1} and O^{D_2} from different sources merged. Formally

$$O =< (K_R^F \cup K_R^{D_1} \cup K_R^{D_2}), (K_T^F \cup K_T^{D_1} \cup K_T^{D_2}), (K_A^F \cup K_A^{D_1} \cup K_A^{D_2}) >$$

In such a case it may be important not only to trace provenance data for assertional axioms (ABox), which form the extensional knowledge in ontologies (or knowledge bases and RDF data collections associated to them), but also provenance data for terminology or role axioms (TBox, RBox), which form the intensional knowledge of the ontology. Such provenance data then describe the origin of each individual terminology/role axiom.

As we need to be able to assign provenance information to all elements of ontology, that means to RBox axioms, TBox axioms and ABox axioms, we need to be able to 'talk' about these axioms. This is the well-known problem already investigated e.g. in [10]. Some of known solutions are presented in [12]: 1) it is possible to use an extensive metamodel of base OWL DL ontology that reifies all its axioms, 2) we can include meta information in annotation properties of the base ontology, 3) it is possible to annotate all axioms of the original ontology with an URI, and to refer to this unique identifier in the meta ontology.

The first approach introduces an extensive meta ontology structure that can be used for our purpose because it exposes axioms of the base ontology as individuals of the meta ontology, but can be computationally difficult. The second approach is relatively simple but it presents provenance as non-logical information outside the (regular) logical semantics of OWL, and therefore is unusable for our needs. The third approach can be used to easily reify axioms in meta ontology without extensive ontological structures required by first approach. Authors of [12] discourage from using this approach because it requires extension of old OWL 1.0 standard in order to assign URIs to axioms. As such extension in XML based syntax they suggest the approach of SWRL, which allows URI references as an optional element [7]. In this paper we however suggest to overcome the drawback of the third approach by relying on the recent OWL 2 standard which allows every axiom of base ontology O to be annotated with a unique identifier – URI [9].

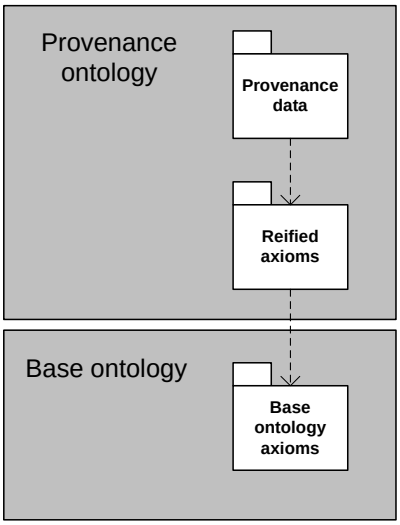

Fig. 1. Ontology Structure

2.3 Provenance Representation Pattern

A general overview of our approach to provenance metadata representation is depicted on Fig. 1. The first level of provenance ontology contains individuals representing reified axioms of the base ontology. These individuals are then assigned actual provenance data.

The ontology pattern itself is depicted on Fig. 2, a detailed example of provenance represenation is then depicted on Fig. 3. We consider a base ontology with three axioms: $\alpha_1 \in K_R$, $\alpha_2 \in K_T$, and $\alpha_3 \in K_A$ (on diagram named `rbox-axiom-alpha1`, `tbox-axiom-alpha2`, and `abox-axiom-alpha3`).

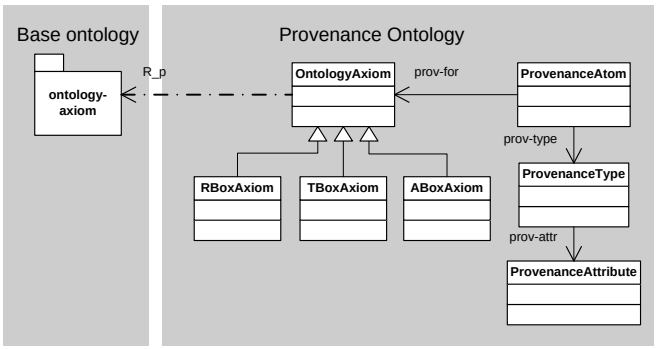

Fig. 2. Provenance pattern

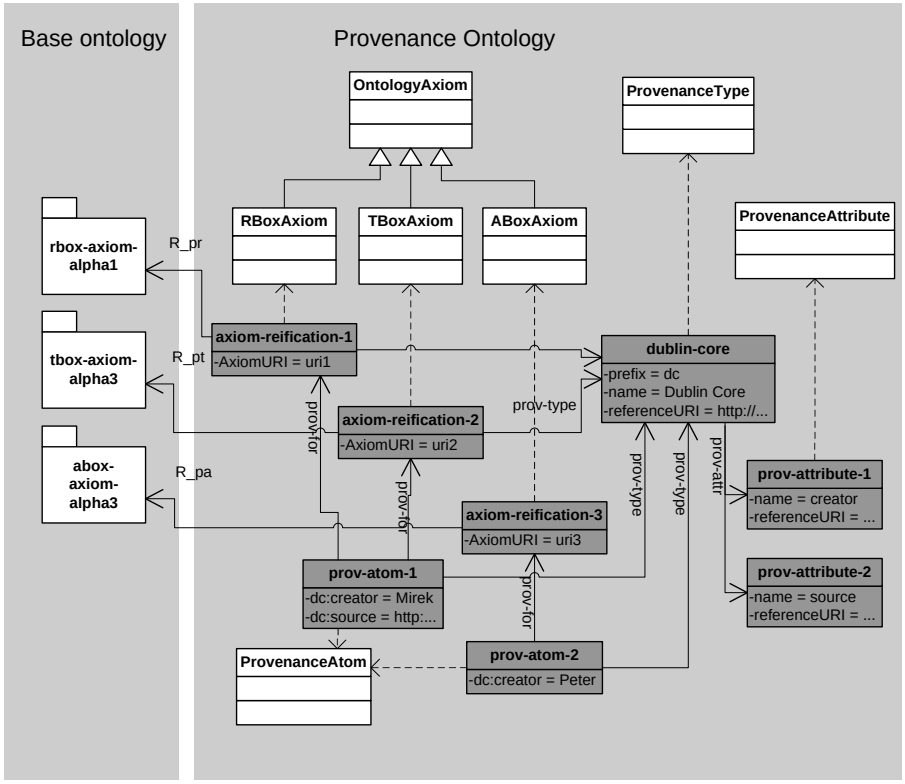

Fig. 3. Provenance pattern example

The description of provenance data is based on the provenance ontology (formally $O^P =< K_R^P, K_T^P, K_A^P >$) that contains reifications of axioms of the base ontology plus the provenance information. The reification level of the provenance ontology consists of class `OntologyAxiom` with subclasses `RBoxAxiom`, `TBoxAxiom`, and `ABoxAxiom`. Individuals belonging to these classes are actually reifications of axioms of base ontology.

We define the reification relation R_{pt} for TBox axioms as follows. Let a be an individual of the provenance ontology and let α be an axiom of the base ontology. Then $R_{pt}(a, \alpha)$ iff α is a TBox axiom and is annotated by a unique identifier `URI` and a is an individual belonging to class `TBoxAxiom` and its data type property `AxiomURI` has value `URI`. We presuppose that R_{pt} is functional and injective.

Analogically we define the reification relation R_{pr} for RBox axioms and the reification relation R_{pa} for ABox axioms. We then also define the general reification relation $R_p = R_{pr} \cup R_{pt} \cup R_{pa}$. Note that reification relations are not DL relations defined in the ontology but *meta-logical* relations. R_p is relation connecting *individuals* of ontology O^P with *axioms* of ontology O.

Reified axioms are then assigned provenance information using the relation `prov-for` to individuals of class `ProvenanceAtom`. Note that the relation `prov-for` is N:N, so a reification of an axiom can be assigned multiple provenance information atoms (i.e. the same axiom was included in multiple original ontologies) and multiple axioms can be assigned a single provenance information atom (ontology from one source has usually multiple axioms). Each individual of this class has defined some provenance information as its datatype properties. It can be for example property "dc:creator" with value "John". Each provenance atom individual is in relation `prov-type` with individuals of class `ProvenanceType`. This class is used to define what kind of provenance definition or standard are we using. Our example on Fig. 3 uses the well-known Dublin Core standard. For provenance types we can define list of attributes that each standard supports by class `ProvenanceAttribute` linked to class `ProvenanceType` by relation `prov-attr`. This approach enables us to use annotations by various provenance meta data standards in single ontology. This is important feature when working with provenance in heterogeneous area of World Wide Web.

3 Propagation of Provenance Metadata in Ontologies

Ontologies do not serve only as static (meta)information representation tool but also enable user to infer new knowledge. Inferred knowledge then can enrich the ontology or it can be used for another purpose. In any case we consider tracking provenance information for inferred knowledge necessary.

Typical DL inferred knowledge in ontologies may include following [2]:

1. $C_1 \sqsubseteq C_2$
2. $C_1 \equiv C_2$
3. $C_1 \cap C_2 = \emptyset$
4. $C \sqsubseteq \bot$ (equivalent to assertion that concept C is unsatisfiable)
5. $C(a)$ (for some arbitrary concept C and individual a).

These are the most common inference tasks results for ontologies, which can be performed by most reasoning engines and also their resulting assertions. It is now necessary to assign this inferred assertions appropriate provenance information. Natural way is to assign provenance meta data to this new knowledge on the basis of provenance meta data of knowledge from which it was inferred.

We denote α an axiom that is inferred from ontology O, therefore $O \vDash \alpha$. Now Kalyanpur et ales. [8] denotes $\mathsf{JUST}(\alpha, O) \subseteq O$ as such fragment of ontology O, that $\mathsf{JUST}(\alpha, O) \vDash \alpha$ and $\forall O'((O' \subset \mathsf{JUST}(\alpha, O)) \rightarrow (O' \nvDash \alpha))$. Informally this set is *justification* for inferred axiom α in ontology O. Next, $\mathsf{ALLJUST}(\alpha, O)$ denotes set of all justifications for α in O, formally $\{O'; O' = \mathsf{JUST}(\alpha, O)\}$ and we define $O_{AJ(\alpha)} = \bigcup \mathsf{ALLJUST}(\alpha, O)$. This is just formal step because while $\mathsf{ALLJUST}(\alpha, O)$ is set of sets of axioms, our defined $O_{AJ(\alpha)}$ is set of axioms of O (formally $O_{AJ(\alpha)} \subseteq O$) what is more appropriate for our use.

When the axiom α is inferred it does not have assigned any provenance information. First it is necessary to annotate (based on OWL 2) the axiom

with new unique URI, so it can be reified on first level of provenance ontology. Then new individual a of class `OntologyAxiom` (and its respective subclass) is introduced to provenance ontology as this reification, with data type property `AxiomURI` having as its value the URI of the axiom α. Formally now $R_p(a, \alpha)$ as we defined earlier.

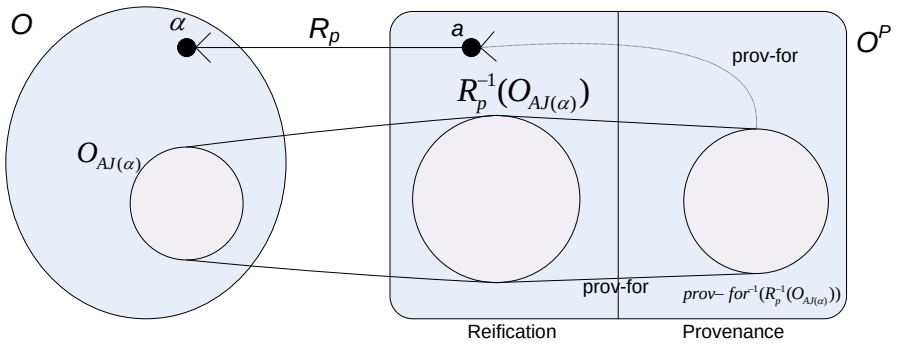

Fig. 4. Provenance propagation

Now we can use set of axioms $O_{AJ(\alpha)}$ and get its respective set of reifications at the first level of provenance ontology using relation R_p. Formally we denote this set of reifications $R_p^{-1}(O_{AJ(\alpha)})$ (see. Fig. 4). These reifications have some provenance information assigned by relation `prov-for` and individual provenance information atoms of class `ProvenanceAtom`. We can formally denote appropriate set of provenance atoms as $\mathsf{prov\text{-}for}^{-1}(R_p^{-1}(O_{AJ(\alpha)}))$. Now this is set of provenance atoms that are assigned to all axioms that are justifications for our inferred axiom α. Thats why we assign this provenance information to this axiom. We know that a is reification of axiom α, so now for every individual x of set $\mathsf{prov\text{-}for}^{-1}(R_p^{-1}(O_{AJ(\alpha)}))$ we add to provenance ontology instance of relation $\mathsf{prov\text{-}for}(x, a)$.

4 Conclusions

Tracking the provenance of application data as well as of ontology elements is one of critical aspects of the Semantic web. We presented an approach to uniformly represent provenance information for data as well as axioms, which relies on a design pattern in OWL. The extended capabilities of the recent OWL 2 version of the language is taken into account. A general method of provenance propagation during ontology-based reasoning is also outlined.

5 Acknowledgments

This work has been partially partially supported by the IGA VSE grant 20/08, IGS 4/2010 and by the CSF grant P202/10/0761 (Web Semantization).

References

1. Richard Arndt, Raphael Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a Well-Founded Multimedia Ontology for the Web. In *Proceedings of The 6th International Semantic Web Conference (ISWC)*, 2007.
2. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, May 2001.
4. Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data Provenance: Some Basic Issues. In *Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 1974. Springer Verlag, 2000.
5. Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and Where: A Characterization of Data Provenance. In *Proceedings of the 8th International Conference on Database Theory*, volume 1973. Springer Verlag, 2001.
6. Juan Gómez-Romero, Fernando Bobillo, and Miguel Delgado. An Ontology Design Pattern for Representing Relevance in OWL. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon J B Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea*, volume 4825 of *LNCS*, pages 71–84, Berlin, Heidelberg, November 2007. Springer Verlag.
7. Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission, World Wide Web Consortium, 2004.
8. Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, and Evren Sirin. Finding All Justifications of OWL DL Entailments. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 267–280. Springer, 2007.
9. Boris Motik, Peter F. Patel-Schneider, and Bijan Parsia. OWL 2 Web Ontology Language, Structural Specification and Functional-Style Syntax. W3C Recommendation 27 October 2009, World Wide Web Consortium, 2009.
10. Steffen Staab and Alexander Maedche. Axioms are objects too: Ontology engineering beyong the modeling of concepts and relations. Research report 399, Institute AIFB, Karlsruhe, 2000, 2000.
11. Miroslav Vacura, Vojtěch Svátek, Pavel Smrž, and Nick Simou. A Pattern-based Framework for Representation of Uncertainty in Ontologies. In *Proceedings of Uncertainty Reasoning for the Semantic Web (URSW)*, 2007.
12. Denny Vrandecic, Johanna Völker, Peter Haase, Duc Thanh Tran, and Philipp Cimiano. A Metamodel for Annotations of Ontology Elements in OWL DL. In York Sure, Saartje Brockmans, and Jürgen Jung, editors, *Proceedings of the 2nd Workshop on Ontologies and Meta-Modeling*, Karlsruhe, Germany, OCT 2006. GI Gesellschaft fur Informatik.