Reflections on Provenance Ontology Encodings

Li Ding[1], Jie Bao[1], James R. Michaelis[1], Jun Zhao[2], and Deborah L. McGuinness[1]

[1] Tetherless World Constellation, Rensselaer Polytechnic Institute
{dingl,baojie,michaj6,dlm}@cs.rpi.edu
[2] Image Bioinformatics Research Group, Department of Zoology, University of Oxford
jun.zhao@zoo.ox.ac.uk

Abstract. As more data (especially scientific data) is digitized and put on the Web, it is desirable to make provenance metadata easy to access, reuse, integrate and reason over. Ontologies can be used to encode expectations and agreements concerning provenance metadata representation and computation. This paper analyzes a selection of popular Semantic Web provenance ontologies such as the Open Provenance Model (OPM), Dublin Core (DC) and the Proof Markup Language (PML). Selected initial findings are reported in this paper: (i) concept coverage analysis – we analyze the coverage, similarities and differences among primitive concepts from different provenance ontologies, based on identified themes; and (ii) concept modeling analysis – we analyze how Semantic Web language features were used to support computational provenance semantics. We expect the outcome of this work to provide guidance for understanding, aligning and evolving existing provenance ontologies.

Keywords: provenance, ontology, semantic web.

1 Introduction

In distributed and open environments, such as the Web, consumers can access data without knowledge of its creation and expected use. Provenance plays an important role in supporting transparency and accountability of such data. In order to ensure data transparency, the corresponding provenance metadata should be made accessible for consumers through an effective environment for data management. Likewise, in order to evaluate the accountability of data, consumers need to correctly understand the provenance metadata. In both cases, the use of provenance ontologies can be helpful. We found it natural to use a Semantic Web-based approach in our work since semantic technologies have been integrated into the web and semantic web languages provide a means for encoding provenance concepts and their meanings.

A number of applications of provenance data management have benefited from the use of Semantic Web ontologies [1, 2, 3]. In these applications, provenance data is represented through RDF graphs serialized in an RDF syntax, such as RDF/XML, and can be published on the Web as Linked Data [4]. Likewise, to encode richer provenance data, either RDFS or OWL are used. Semantic Web tools, such as RDF APIs, triple stores and reasoners, are used to leverage the vocabulary defined by the provenance ontologies and make inferences accordingly. In particular, SPARQL can be used to express simple queries over provenance data.

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 198–205, 2010.

We selected a few representative Semantic Web provenance ontologies for analysis, and attempt to address two issues: (i) **Concept coverage**: what primitive provenance concepts are supported? What are the similarities and differences among the primitive concepts? (ii) **Concept modeling**: how are computational provenance semantics modeled? What are the expressivities? The outcome of our study not only provides guidance to provenance ontology users but also has the potential to promote best practices in collaborative provenance ontology development.

The rest of this paper is organized as follows: Section 2 reviews selected Semantic Web provenance ontologies; Section 3 analyzes primitive provenance concepts in these ontologies using theme-based grouping; Section 4 analyzes the use of ontology constructs in provenance ontologies; Finally, Section 5 provides concluding remarks.

2 Semantic Web Provenance Ontologies

Semantic Web provenance ontologies[1] have emerged in one of two ways: from scratch, or through the conversion of existing provenance vocabularies not based on Semantic Web technologies. Many of them focus on the provenance of digital objects (e.g. scientific data), using the Web as a data management infrastructure, and encoding computational provenance semantics using declarative ontology constructs. This paper focuses on the following representative ontologies[2].

Open Provenance Model (OPM) [5] originated from workflow trace sharing, and is also designed to support more general provenance representation and computation. It defines both provenance *entities* (e.g., "artifact") and provenance *relations* (e.g., an artifact "was generated by" a process). Starting from XML Schema-based encodings, OPM recently adopted an OWL-based encoding, which has evolved with the new OPM specification (OPM 1.1).

Proof Markup Language (PML) [6] originated from logical proof sharing, and has also been used in other information manipulation contexts such as information extraction and machine learning. It consists of three modules: (i) a taxonomy of primitive provenance entities with related properties; (ii) a representation of data derivation and acquisition trace using proof theoretic notation; and (iii) an encoding of trust and belief on data and agents. Since its inception in 2003, PML has consistently followed the linked data principle in publishing its data.

Dublin Core (DC) was originally developed in the digital library domain. It provides a provenance vocabulary which primarily covers generic binary provenance relations such as "source" and "creator". It leverages RDFS ontology constructs, and its provenance relations are typically binary without specifying domain/range restrictions. **Dublin Core Terms (DCTerms)** [3] is the current recommended version of Dublin Core, having more relations and concepts than the previous DC Element vocabulary[4].

Provenance Vocabulary (PRV) [7] was recently developed to track information manipulation. It consists of three modules: (i) the core module that defines basic concepts for tracking data creation and data access, (ii) a taxonomy specific to Web

[1] In the rest of this paper, the term "ontology" refers to semantic web provenance ontology.
[2] For more provenance related ontologies, see http://tw.rpi.edu/portal/Provenance.
[3] http://dublincore.org/documents/dcmi-terms/
[4] http://dublincore.org/documents/dces/

information transfer and (iii) a taxonomy specific to authentication of information. It uses OWL 2 constructs, e.g., property chain Axiom.

Provenir[5] [14] focuses on information manipulation. It is built on top of the OBO Relation Ontology (OBO-RO) [8], which covers generic binary relations frequently used in bioinformatics. This ontology defines new provenance entities, and it also uses the newly defined classes to extend the definition of OBO-RO by adding domain and range restrictions to existing OBO-RO properties.

Some of the above ontologies are modularized. Table 1 lists the selected ontologies as well as their key modules (with the corresponding namespace-prefix mappings). Every module can be retrieved by dereferencing the corresponding namespaces, except OPM[6].

Table 1. Selected semantic web provenance ontologies and their modules

Ontology	Namespace	Prefix
OPM 1.1	http://openprovenance.org/ontology#	opm
PML 2.0	http://inference-web.org/2.0/pml-provenance.owl# http://inference-web.org/2.0/pml-justification.owl#	pmlp pmlj
Dublin Core Terms	http://purl.org/dc/terms/	dcterms
Provenance Vocabulary	http://purl.org/net/provenance/ns#	prv
Provenir Ontology OBO Relation Ontology	http://knoesis.wright.edu/provenir/provenir.owl# http://www.obofoundry.org/ro/ro.owl#	provenir ro

Table 2 shows basic statistics about these ontologies: triples are counted using the JENA API[7]; class/property numbers are counted based on two criteria, i.e. (i) the terms are *defined as classes or properties* [9] and (ii) the terms use the module's namespace (we did not count redefined external concepts in PRV and Provenir); and OWL species and DL Expressivity were obtained using Pellet[8] online services.

Table 2. Basic statistics of selected semantic web provenance ontologies

	opm	pmlp	pmlj	dcterms	Prv	provenir	ro
# of triples	309	505	207	857	304	136	268
# of classes	20	30	8	22	14	8	0
# of properties	26	47	21	55	17	2	24
OWL Species	OWL DL	OWL DL	OWL DL	RDFS	OWL 2 DL	OWL DL	OWL Lite
DL Expressivity	ALCF(D)	ALCHIF(D)	ALHF(D)	ALH(D)	RI(D)	ALCH	$AL_{R+}HI$

3 Concept Coverage Analyses

We review concept coverage from two perspectives: (i) to group similar primitive concepts by their themes to see if different ontologies focus on similar themes; and (ii) to review semantics of primitive concepts for identifying the difference between similar

[5] http://wiki.knoesis.org/index.php/Provenir_Ontology

[6] http://github.com/lucmoreau/OpenProvenanceModel/raw/master/elmo/src/main/resources/opm.owl was used in this study as the current draft OWL profile for OPM 1.1 .

[7] http://jena.sourceforge.net/

[8] Pellet online demo at http://www.mindswap.org/2003/pellet/demo.shtml

primitive concepts. Due to their different design principles, primitive concepts in different ontologies are not necessarily the same even if they have the same name. Since the meaning of primitive concepts is primarily described in natural language in the annotations of ontology and the related publications, precise alignment is challenging. Therefore, we empirically identified several themes to use for grouping similar primitive concepts and further discuss their differences. We are not claiming comprehensive coverage with our provenance concept themes but we did find them instructive for provenance ontology comparison.

Table 3 lists the selected ontologies and compares their concept theme coverage. Due to space limitations, each table cell contains a few example terms from the corresponding ontology (see rows) on the theme (see columns). To support definition of themes, we use "entity" to refer to things that distinctly exist and "relation" to refer to relations among entities. A "theme" is used to group similar entities and relations reflecting one dimension of provenance metadata, and these themes have clear connection to the well-known five Ws (and one H)[9] in information gathering. Themes are identified based on an empirical study over the selected provenance ontologies[10], so it is not necessarily exhaustive. Themes are generally disjoint, but exceptions are permitted. For example, a robot could be considered as an agent in a car manufacturer factory, but an artifact (product) of a robot manufacturing factory.

Table 3. Provenance ontology theme coverage

	OPM 1.1	PML 2.0	DCTerms	PRV core	Provenir (+OBO-RO)
Agents	Agent	Agent	Agent	Actor	Agent
artifacts	Artifact	IdentifiedThing,Information	PhysicalResource	Artifact	Data
Events	WasGeneratedBy Process	pmlp:SourceUsage, pmlj:InferenceStep	ProvenanceStatement Source	Execution	provenir:process, ro:derives_from
methods		InferenceRule	Policy MethodOfAccrual	CreationGuideline	
time	OTime	hasCreationDateTime	PeriodOfTime	performedAt	temporal_parameter
space			Location		Spatial_parameter

- **Agents (Who):** Actionable entities that can take actions in an event. Organization and Person are two common types of agents. PML and PRV core additionally defined agent taxonomies. While opm:Agent is defined as a snapshot of an agent, the others define Agent as a continuant entity which is mutable over time.
- **Artifacts (Who):** Entities made by agents and involved in events. OPM explicitly emphasizes the immutable status of artifacts, such that an evolving entity could be related to multiple artifacts (each of which being a snapshot of the entity). While OPM and DCTerms consider both digital and physical entities, the other selected ontologies focus only on digital entities, especially data. opm:Artifact is defined to be disjoint with opm:Agent. Both PML and PRV additionally define artifact taxonomies. PML defines pmlp:Information (i.e. snapshot of data) and pmlp:Source (i.e. the container of information) to support differentiating inference steps (i.e., deriving data from data) from source usage (e.g., acquiring data from data containers).

[9] http://en.wikipedia.org/wiki/Five_Ws (Who, What, When, Where, Why, & How)

[10] Note: our study is based on the ontology and supporting documents. We are in discussion with ontology authors to confirm and refine our observations.

- **Events (What):** Observable occurrence(s), execution of action(s) (potentially including the past). Although not explicitly claimed, these ontologies contain entities and/or relations that record events, especially derivation, i.e., something was derived from something else. For example, opm:WasGeneratedBy captures an event where an artifact was generated by a process at a certain moment, and opm:Process is also a kind of event because "*processes also occurred in the past*"[5]. PML supports both data derivation events by pmlj:InferenceStep and data acquisition events by pmlp:SourceUsage. OBO-RO and DCTerms offer binary relations, e.g., ro:derives_from and dcterms:source, which obviously can be mapped to data derivation events and data acquisition events, respectively.
- **Methods (How):** Entities denoting the operations (or actions) used (or mentioned) in events. For example, a recipe exposes the instructions used in a cooking event, and a protocol shows the methods used in a biomedical experiment. PML uses pmlp:InferenceRule to annotate methods used in events so that users can find events reusing the same method. DCTerms and PRV also have similar concepts. An instance of method can be further declaratively annotated by declarative scripts such as list of instructions or program source code.
- **Time (When):** Temporal concepts, such as time and date when things were created (or updated), primarily used for annotating events. Most ontologies only defined temporal properties, while OPM and DCTerms define additional time classes. Unlike the other selected ontologies which only focus on time points, OPM additionally defines duration using *opm:noEarlierThan* and *opm:noLaterThan.*
- **Space (Where):** Geospatial concepts such as locations, GPS coordinates and regions. Only DCTerms and Provenir support this theme, and their definitions are remain general and avoid including detailed geospatial concept taxonomies.

A few additional observations arose with the theme analysis. First, similar concepts may still have different meanings, e.g., OPM and PML treat the concept "agent" as immutable and mutable, respectively. Second, feedback from the use of provenance ontologies in applications can lead to their evolution, e.g., a special concept pmlp:LearnedSourceUsage was added to better support explaining tasks in multi-agent learning contexts[10]. Third, some themes can be supported by dedicated ontologies, e.g., the OWL Time ontology[11], the WGS84 Geo Positioning Ontology[12] and Friend of a Friend (FOAF)[13]. Finally, it is important to represent the scope of a particular workflow, e.g., OPM defined opm:Account to associate entities with a workflow, and PML uses a recursive algorithm to determine the scope of a proof.

4 Concept Modeling Analyses

We now analyze the semantic structure and concept modeling patterns by comparing the use of ontology constructs in the ontologies (see Table 4). The ontology constructs are further grouped by the following four functional groups that were summarized from the manual analysis of the selected ontologies.

[11] http://www.w3.org/2006/time
[12] http://www.w3.org/2003/01/geo/wgs84_pos
[13] http://xmlns.com/foaf/0.1/

Table 4. How OWL/RDFS ontology constructs were used in provenance ontologies

		opm	pmlp	pmlj	dcterms	prv	provenir	ro
Concept Taxonomy	rdfs:subClassOf	X	X	X	X	X	X	
	rdfs:subPropertyOf		X	X	X	X	X	X
	owl:disjointWith	X	X			X	X	
	owl:unionOf	X				X	X	
	owl:equivalentClassOf					X		
Inference on relations	owl:inverseOf		X			X		X
	owl:TransitiveProperty							X
Constraints	rdfs:domain / rdfs:range	X	X	X	X	X	X	
	owl:allValuesFrom	X	X	X			X	
	Cardinality Restriction	X	X	X		X		
Concept Reuse	owl:imports		X	X				
	Reused ontology				foaf		ro	

Concept Taxonomy. Semantic ontology languages, e.g., RDFS and OWL, provide set-theoretic constructs (e.g. sub-set, union, equivalence, complement and disjoint) to support taxonomy definitions. These ontology constructs are observed in all selected ontologies in modeling class taxonomies and/or property taxonomies. A direct benefit of using OWL and RDFS is that those constructs are supported by corresponding reasoners that are capable of inferring additional information about taxonomies, e.g., transitive closure of sub-set relations and consistency validation using disjoint semantics. PML has a larger class taxonomy than some other provenance interlingua options (e.g. OPM). One reason for PML's growth was a direct consequence of application driven growth from reuse beyond its original scope (e.g., reuse in machine learning and text analytics applications although original constructor design was aimed at hybrid logical first order reasoning. Its growth also generated a redesign to create modules that could be used independently. We anticipate that other provenance ontologies, if they decide to grow in breadth, may also provide modularization options. The use of disjointness may be an issue in some ontologies. Overusing disjointness can limit reusability since for example an initial modeling might expect person and inference-engine to be disjoint but in a different context, a person might function as an inference engine and thus may be an instance of both classes.

Inference on Relations. A binary provenance relation can be defined as an OWL object property to carry additional computational semantics such as "inverse" and "transitive". Upon defining "part/whole" relation, OBO-RO additionally used both owl:TransitiveProperty and owl:inverseOf constructs in defining ro:has_part in comparison with a similar concept dcterms:hasPart from DCTerms. Besides the two constructs in table 4, PRV leverages the OWL2 construct owl:propertyChainAxiom to enable more complex inference on relations: if x is prv:serializedBy y and y is prv:createdBy z then x is prv:createdBy z. The OWL and OWL2 constructs discussed here are selected because they have obvious connection to provenance graph inference [5]. We should also note that SPARQL can also be used to enable some other kinds of provenance computation, such as converting binary relation from/to corresponding class instances.

Constraints. Upon sharing provenance metadata, users may also want to leverage provenance ontologies to assure the quality of provenance metadata. Integrity constraints,

such as cardinality restrictions, may be encoded using the OWL syntax along with a non-standard semantics based on the Closed World Assumption (CWA) [11][14]. For example, an instance of opm:WasGeneratedBy needs to be associated with at least one instance of opm:Process via the opm:cause relation, and this can be encoded using owl:minCardinality.

Concept Reuse. Section 2 showed the trend of modularizing provenance ontologies, and raised issues on ontology reuse. In practice, ontology reuse can be done by using the owl:imports construct (i.e., explicitly copy the content of the other ontology, e.g., pmlj imports pmlp), or by directly using terms in the external ontology (i.e., users need to dereference the terms to get their definitions, e.g., PRV uses external terms to enrich its definition). It is notable that the meaning of imported terms may also be redefined during importing. For example, in Provenir, the meaning of ro:has_agent is beyond its original meaning in OBO-RO due to additional domain/range statements.

5 Conclusion

This study investigated a select group of Semantic Web provenance ontologies and yielded interesting observations: (i) provenance ontologies share common themes surrounding provenance research; (ii) similar terms in the same theme can carry different semantics, e.g. opm:Agent and pmlp:Agent; (iii) we observed the use of RDFS, OWL and OWL2 in encoding provenance ontologies and supporting provenance computation (e.g. transitive provenance graph inference); (iv) Some ontologies are fully self-contained while some others reuse external concepts.

The above observations not only help users to review provenance ontologies via a side-by-side comparison, but also promote better collaborative provenance ontology development. First, modularization has been seen as a successful practice in ontology development for controlling the cost of development and reuse. It would be desirable to keep a minimal set of core concepts in one module and support extensions, such as detailed classification and domain specific concepts, in other modules. However, we should also avoid excessive modularization that may cause unnecessary overhead. Second, while provenance theme-level mapping provides general guidance for reusing ontologies, a concept-level mapping is still needed to keep the ontologies interoperable. Work on ontology mapping has been reported on the OPM-PML mapping [12] and the OPM-DCTerms mapping [13]. Additionally, mapping efforts are underway in the W3C Provenance Incubator[15]. Future research should also emphasize mapping different provenance ontologies as well as reusing concepts from other ontologies.

Acknowledgments. This work is supported in part by NSF #0524481, DARPA #FA8650-06-C-7605, #FA8750-07-D-0185, #55-002001, #F30602-00-2-0579, and ITA project W911NF-06-3-0001.

[14] Note that the standard semantics of OWL does not support the modeling of integrity constraints as it uses the Open World Assumption (OWA), c.f. [11].

[15] http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings

References

[1] Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M.: Using semantic web technologies for representing e-science provenance. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 93–106. Springer, Heidelberg (2004)

[2] Golbeck, J., Hendler, J.: A Semantic Web approach to the provenance challenge. Concurrency and Computation: Practice and Experience 20, 431–439 (2008)

[3] Zednik, S., Fox, P., McGuinness, D.L., Pinheiro da Silva, P., Chang, C.: Semantic Provenance for Science Data Products: Application to Image Data Processing. In: Workshop on the role of Semantic Web in Provenance Management (2009)

[4] McGuinness, D.L., Pinheiro da Silva, P.: Explaining Answers from the Semantic Web: The Inference Web Approach. Journal of Web Semantics 1(4), 397–413 (2004)

[5] Moreau, L., Clifford, B., Freire, J., Gil, Y., Groth, P., Futrelle, J., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Simmhan, Y., Stephan, E., Bussche, J.: The Open Provenance Model Core Specification (v1.1), submitted to Future Generation Computer Systems (2009)

[6] McGuinness, D.L., Ding, L., Pinheiro da Silva, P., Chang, C.: PML 2: A Modular Explanation Interlingua. In: Proceedings of the 2007 Workshop on Explanation-aware Computing, ExaCt-2007 (2007)

[7] Hartig, O., Zhao, J.: Publishing and Consuming Provenance Metadata on the Web of Linked Data. In: Proceedings of the 3rd International Provenance and Annotation Workshop, IPAW (2010)

[8] Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in Biomedical Ontologies. Genome Biology, 6:R46 (2005)

[9] Ding, L., Finin, T.: Characterizing the Semantic Web on the Web. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 242–257. Springer, Heidelberg (2006)

[10] McGuinness, D.L., Glass, A., Wolverton, M., Pinheiro da Silva, P.: Explaining Task Processing in Cognitive Assistants that Learn. In: AAAI 2007 Spring Symposium on Interaction Challenges for Intelligent Assistants (2007)

[11] Tao, J., Sirin, E., Bao, J., McGuinness, D.L.: Integrity Constraints in OWL, accepted by the 24th Conference on Artificial Intelligence (AAAI 2010)

[12] Michaelis, J.R., Zednik, S., Ding, L., McGuinness, D.L.: A comparison of the OPM and PML provenance models. Tetherless World Constellation Technical Report (2009)

[13] Miles, S., Moreau, L., Futrelle, J.: OPM Profile for Dublin Core Terms (v0.3) (July 2009), `http://twiki.ipaw.info/pub/OPM/ChangeProposalDublinCoreMapping/dcprofile.pdf`

[14] Sahoo, S.S., Barga, R.S., Goldstein, J., Sheth, A.: Provenance Algebra and Materialized View-based Provenance Management, Microsoft Research Technical Report (MSR-TR-2008-170) (November 2008)