Provenir ontology: Towards a Framework for eScience Provenance Management

Satya S. Sahoo, Amit P. Sheth
Kno.e.sis center, Computer Science and Engineering Department, Wright State University, Dayton, OH-45324, USA
{sahoo.2, amit.sheth}@wright.edu

Abstract

Provenance metadata describes the “lineage” or history of an entity and necessary information to verify the quality of data, validate experiment protocols, and associate trust value with scientific results. eScience projects generate data and the associated provenance metadata in a distributed environment (such as myGrid) and on a very large scale that often precludes manual analysis. Given this scenario, provenance information should be, (a) interoperable across projects, research groups, and application domains, and (b) support analysis over large datasets using reasoning to discover implicit information. In this paper, we introduce an ontology-driven framework for eScience provenance management underpinned by an “upper-level” ontology called *provenir* defined in OWL-DL. This framework is implemented in a modular fashion by extending *provenir* ontology to create a suite of domain-specific provenance ontologies that facilitate interoperability and enable reasoning. We demonstrate the application of this framework in two eScience projects domains through creation of, (a) Parasite Experiment ontology to model provenance in parasite research, and (b) Trident ontology to model provenance in the Neptune oceanography project.

Introduction

Provenance, from the French word “provenir” meaning “to come from”, describes the lineage or history of an entity. Provenance metadata in eScience is necessary to accurately interpret data, compute trust value associated with scientific results, and ensure correct use of data. Provenance in eScience projects is generated in a distributed environment, using potentially heterogeneous experiment methods; hence interoperability of provenance information is essential to allow effective comparison and/or integration of the scientific data. Further, provenance metadata generated in high-throughput experiments can be analyzed effectively by software applications that use complex inference rules to discover implicit information. In this paper, we describe an framework for provenance management underpinned by an upper level provenance ontology called *provenir* and demonstrate its application in two eScience projects.

Provenir upper-level provenance ontology

The *provenir* ontology is based on our earlier work that led to the creation of ProPreO ontology, a provenance ontology for proteomics [1]. *Provenir* ontology is defined in OWL-DL, OWL-DL [2] represents the most expressive but decidable sub-language of the W3C Web Ontology Language (OWL). The *provenir* ontology defines three base classes representing the primary components of provenance, that is, “`data`”, “`agent`” and “`process`”[1] (Figure 1). The datasets that undergo modification in an experiment are modeled as `data_collection` class and the parameters that influence the execution of experiments are modeled as `parameter` class. Both these classes are

[1] We use the `courier` font to denote ontology classes and relationships

sub-classes of the data class. The parameter class has three sub-classes representing the spatial, temporal and thematic (domain-specific) dimensions, namely spatial_parameter, temporal_parameter, and domain_parameter. Instead of defining a new properties, a set of 11 fundamental properties defined in the Relation ontology (RO) [3] have been adapted and defined in terms to *provenir* ontology classes [4].

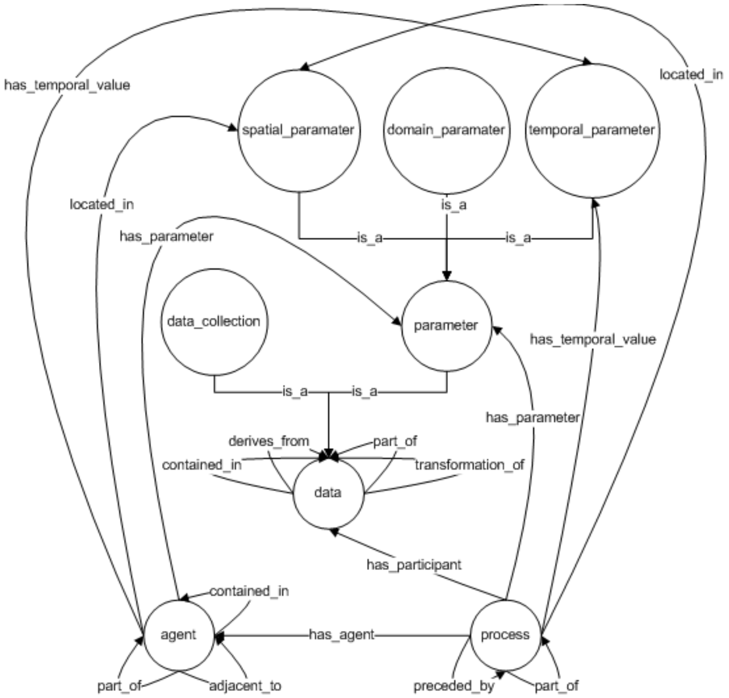

Figure 1: *Provenir* ontology schema

In contrast to the Open Provenance Model (OPM) [5], a similar effort to create a common model for representation of provenance, *provenir* ontology is more expressive both in terms of the modeled concepts and well-defined named relationships. This enables *provenir* ontology to be easily extended for modeling of complex domain-specific provenance information that is difficult or not possible in OPM. Further, *provenir* supports complex provenance analysis using the extensive Semantic Web reasoning framework (SWRL and now RIF) [6], while inference in OPM is limited and error-prone [7] due to its generic graph structure.

Domain-specific information or “domain semantics” is an important aspect of provenance in eScience. But, a single monolithic provenance ontology to model details from different domains is clearly not feasible. Hence, our provenance framework involves integrated use of multiple ontologies, each modeling provenance metadata specific to a particular domain. The use of *provenir* as the upper-level reference ontology facilitates interoperability across the domain-specific provenance ontologies.

Parasite Experiment ontology

The Parasite Experiment (PE) ontology was developed as part of the NIH-funded *T.cruzi* Semantic Problem Solving Environment (SPSE) project [8]. The PE ontology extends the classes and relationships in *provenir* ontology to model provenance information associated with “Gene Knockout” (GKO) and “Strain Creation” (SC) experiment protocols (Figure 2). The GKO and SC protocols consist of multiple sub-processes, which are modeled in PE ontology as sequence_extraction, plasmid_construction, transfection, drug_selection, and cell_cloning classes (Figure 2).

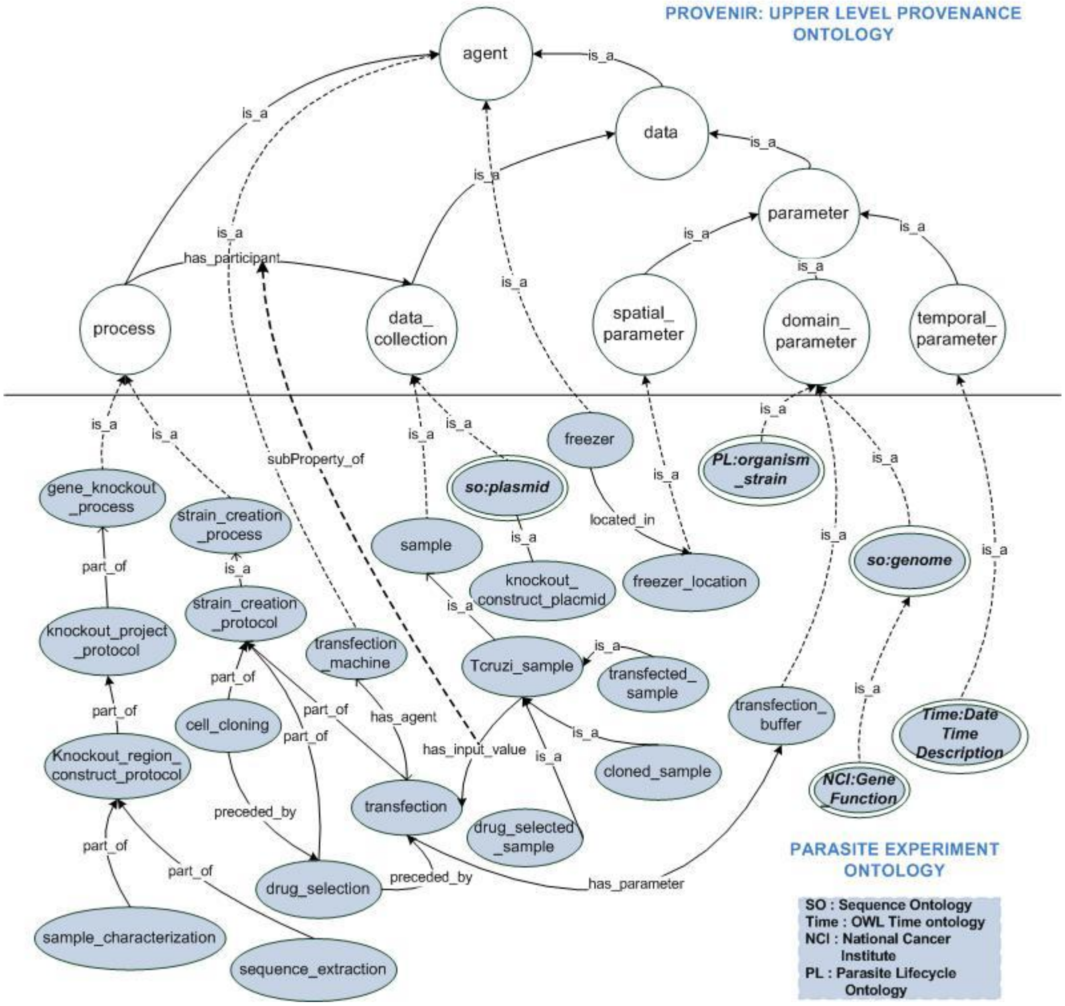

Figure 2: Parasite Experiment ontology

The data entities and parameter used in the experiment protocols are also modeled, for example, given the transfection process its input value Tcruzi_sample is modeled as specialization of provenir:data_collection class, whereas the parameter value transfection_buffer is modeled as specialization of the provenir:parameter class. The PE ontology also models the different types of agents used in parasite research experiments, for example, transfection_machine, microarray_plate_reader are instruments, researcher is an example of human agent; and knockout_plasmid is an example of a biological agent. The PE ontology is modeled using the OWL-DL language and contains 88 classes and 23 named relations. PE ontology is open sourced through the National Center for Biomedical Ontologies (NCBO)[2].

[2] http://bioportal.bioontology.org/ontologies/40425

Trident ontology

The Neptune project [9], led by the University of Washington, is an ongoing initiative to create network of instruments widely distributed across, above, and below the seafloor in the northeast Pacific Ocean. We consider a simulated scenario, illustrated in Figure 3, involving collection of data by ocean buoys (containing a temperature sensor and an ocean current sensor), which is then processed by a scientific workflow.

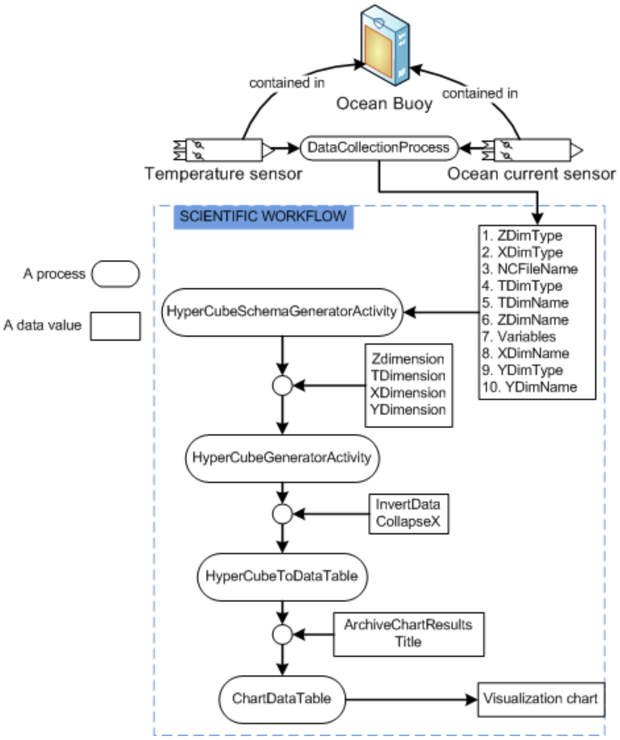

Figure 3: Neptune oceanography scenario modeled in Trident ontology

The scientific workflow is composed of four steps to process the data from the sensors and create visualization charts as output. The Trident ontology models the details of this scenario by extending the *provenir* classes, for example `temperature_sensor` and `ocean_current_sensor` are modeled as specialization of `provenir:agent`. Similarly, the `provenir:spatial_parameter` is extended to model the geographical location (latitude-longitude) of the `ocean_buoy` and `provenir:temporal_parameter` is extended to model date and time details associated with sensor data.

In the next section, we briefly describe the infrastructure created to support the provenance query and analysis over provenance information represented by using domain-specific ontologies such as Trident and PE ontology.

Provenance Management Infrastructure

In addition to *provenir* ontology, a set of specialized provenance query operators have been proposed, as part of the provenance management framework, to support query and analysis of provenance information. The set of query operators are:

(a) **provenance ( )** – to retrieve provenance information for a given dataset,

(b) **provenance_context ( ) –** to retrieve datasets that satisfy constraints on provenance information,

(c) **provenance_compare ( )** – given two datasets, this query operator determines if they were generated under equivalent conditions by comparing the associated provenance information, and

(d) **provenance_merge ( )** – to merge provenance information from different stages of an experiment protocol.

The query operators are defined in terms of the provenir ontology class and relations (formal definition of query operators are presented in [10]). Using standard Resource Description Framework Schema (RDFS) entailment rules, such as subsumption, along with user-defined rules (incorporating domain-specific information), the provenance query operators support queries over any domain-specific provenance ontology that extends *provenir* ontology. A provenance query engine has been implemented over an Oracle RDF store [11] to support the provenance query operators [10].

Conclusion

In this paper, we describe the implementation of an ontology-driven framework for provenance management in eScience projects. The framework consists of an upper-level ontology called *provenir* that can be extended to model interoperable, domain-specific provenance ontologies. The application of the framework is demonstrated in two eScience projects for parasite research and oceanography.

Acknowledgement

We acknowledge the significant help of Roger Barga, Microsoft Research eScience Group, in creation of the Trident ontology. This work was partly funded by NIH Grant#1R01HL087795-01A1.

Reference

[1] Sahoo SS, Thomas, C., Sheth, A., York, W. S., and Tartir, S. Knowledge modeling and its application in life sciences: a tale of two ontologies. In: Proceedings of the 15th international Conference on World Wide Web WWW '06 2006 May 23 - 26; Edinburgh, Scotland; 2006. p. 317-326.

[2] http://www.w3.org/TR/owl-features/. 22 Jan 2008

[3] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46.

[4] Sahoo SS, Barga, R.S., Goldstein, J., Sheth, A.P., Thirunarayan, K. "Where did you come from...Where did you go?" An Algebra and RDF Query Engine for Provenance Kno.e.sis Center, Wright State University; 2009.

[5] http://twiki.ipaw.info/bin/view/Challenge/OPM.

[6] Boley H, Hallmark, G., Kifer, M., Paschke, A., Polleres, A., Reynolds, D. RIF Core Dialect; 2009.

[7] Simmhan YL. FeedbackonOPM. In; 2008.

[8] Sahoo SS, Weatherly, D.B., Muttharaju, R., Anantharam, P., Sheth, A., Tarleton, R.L. Ontology-driven Provenance Management in eScience: An Application in Parasite Research. In: R. Meersman TDea, editor. The 8th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 09); 2009; Vilamoura, Algarve-Portugal: Springer Verlag; 2009.

[9] http://www.neptune.washington.edu/.

[10] Sahoo SS, Barga, R.S., Goldstein, J., Sheth, A.P., Thirunarayan, K. "Where did you come from...Where did you go?" An Algebra and RDF Query Engine for Provenance Kno.e.sis Center, Wright State University; 2009.

[11] Chong EI, Das, S., Eadon, G., and Srinivasan, J. An efficient SQL-based RDF querying scheme. In: 31st international Conference on Very Large Data Bases; 2005 August 30 - September 02; Trondheim, Norway: VLDB Endowment; 2005. p. 1216-1227