A TripleStore implementation of the IVOA Provenance Data model

Mireille Louys[1], Pineau François-Xavier[2], Bonnarel François[2] and Holzmann Lucas[3]

[1]*Icube and CDS, Observatoire astronomique , University of Strasbourg, France;* `mireille.louys@unistra.fr`

[2] *CDS, Observatoire astronomique, University of Strasbourg, France*

[3] *ENSIIE, EVRY, France*

Abstract. The IVOA has proposed a standard for capturing the provenance metadata in the production and distribution of astronomical data. We present an implementation in a triplestore for the provenance information recorded for a collection of astronomical images. The ontology applied is derived from PROV-O from the W3C and from the IVOA Provenance data model. SPARQL queries based on the data model concepts allow to select datasets on a wide range of provenance properties and have proven to be efficient in the triplestore representation. The data model of the SIMBAD CDS data base has also been tested, and turned out to scale very efficiently in the triplestore strategy as well.

1. Goal

This paper presents an evaluation of a triplestore implementation and its comparison with a relational database implementation. The two databases serve provenance metadata as modeled following the IVOA Provenance data model available at `http://www.ivoa.net/documents/ProvenanceDM/`. We present how the data model has been translated to an equivalent ontology and how table data from a Prov-TAP service can be translated to a list of triples following the semantics of the ontology. Equivalent queries have been addressed to both database architectures in order to test expressivity and extensibility as well as completeness for both systems. This exercise is also an experiment on the way the data model classes and relations are put in practice and activated with real datasets. It helped to understand how relations, roles and properties can be enhanced in the data model implementation.

2. Provenance Metadata representation for an astronomical image data base

The IVOA Provenance data model is an IVOA specification for structuring and describing metadata about the history of data sets preparation and publication in astronomy. It represents the operations applied to data the agent performing or responsible for these operations and the entities, typically data sets and parameters involved in these applications.

3. Implementations of the IVOAProvenance DM

The typical implementation of the IVOA provenance DM is through a relational database architecture as done for the RAVE use case, CTA pipe project, CDS test image database and planned for the SVOM project. In this design, each class of the data model is stored as one database table, with each row representing an instance (e.g an Entity instance representing a data set is stored in the Entity table, an agent instance in the Agent table, etc.). Relations between instances are stored as instances (rows) in the corresponding relation table (e.g. "Entity E1 wasGeneratedBy Activity A1" is stored as a row in the wasGeneratedBy table and binds both identifiers from E1 and A1 .)

Another way to represent those metadata is to use RDF/ttl in a triplestore representation, gathering all properties we know about the instances in sentences built on an ontology that translates the relations and attributes defined in the model. As the IVOA Provenance DM extends the W3C PROV-DM, we extended the W3C PROV-O ontology ((Belhajjame et al. 2013)) to tackle the extended part of the model : description, configuration with parameter, etc. The draft ontology is available at `http://wiki.ivoa.net/internal/IVOA/ProvenanceRFC/provOntologyIVOA2018-v1-0.owl` and sketched in Fig. 1.

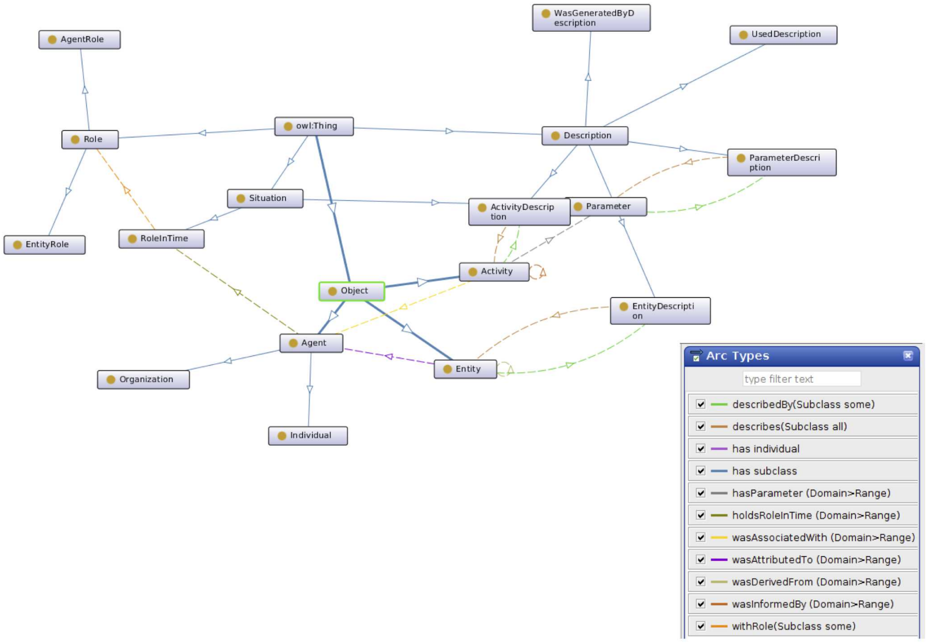
WasGeneratedByDescription
UsedDescription
AgentRole
Role
owl:Thing
Description
ParameterDescription
Situation
ActivityDescription
Parameter
EntityRole
RoleInTime
Activity
Object
EntityDescription
Agent
Entity
Organization
Individual
Arc Types
type filter text
☑ describedBy(Subclass some)
☑ describes(Subclass all)
☑ has individual
☑ has subclass
☑ hasParameter (Domain>Range)
☑ holdsRoleInTime (Domain>Range)
☑ wasAssociatedWith (Domain>Range)
☑ wasAttributedTo (Domain>Range)
☑ wasDerivedFrom (Domain>Range)
☑ wasInformedBy (Domain>Range)
☑ withRole(Subclass some)

Figure 1. Provenance Ontology extending the W3C PROV-O for IVOA Provenance DM

/

3.1. The CDS test image database

We have gathered a subset of images processed for feeding the Aladin image data base server in order to setup a prototype data base for tracing provenance metadata from image digitization to RGB image combination, as well as computation of HIPS datasets. The CDS test database is available via a TAP service, following the IVOA TAP protocol ((Dowler et al. 2010)) built-up on a Postgres data base back-end. It implements the following set of classes and relations :

Classes	**Relations**
Activity	hadDescription
Entity	used wasGeneratedBy wasDerivedFrom
Agent	wasAttributedTo wasAssociatedTo
ActivityDescription	hadDescription
Parameter	hadConfiguration
ParameterDescription	
UsedDescription	
WasGeneratedByDescription	

Table 1. Provenance Metadata supported in the CDS implementation

3.2. Triplestore implementation for CDS test data

The BlazeGraph (`https:/www.blazegraph.com`) platform was chosen as a testbed in order to evaluate the merits of a triplestore for the representation of our CDS test database. All classes and relations have been translated from PostGres tables to VOTable then in CSV individual list, one for each class or relation. The ingestion in the triplestore builds up triples for each instance of a class or of a relation. Every relation is translated in a unique predicate. Roles for Agents need to be created with an extra predicate (holdsRoleInTime) in order to allow one Agent instance to play different roles with respect to an Activity or an Entity.

At the end of the translation, the triplestore entails the same provenance metadata as the CDS prototype PostGres data base. It provides an equivalent representation of provenance information.

The Blazegraph CDS repository is available at `http://cdsgit...tobecompleted`

3.3. Triplestore queries for evaluation on CDS test data

A list of test queries has been defined in order to evaluate the capabilities of the triplestore queries to match with those of the Postgres/TAP service. This is listed on the Provenance DM RFC page at
`http://wiki.ivoa.net/internal/IVOA/ProvenanceRFC/ProvQuerytest-3store.pdf`.
All queries addressed to the TAP service have a translated version in the triplestore implementation and provide a similar list of hits. The equivalence of representation is confirmed and the triplestore offers at least the same expressivity for queries. The query mechanism uses name matching for relations and for class attributes. It filters

the triples on the discovered values . No complex nor intricate joins are required. More investigation is needed to check how the filtering can be ordered to optimize the search.

4. Lessons Learned

The triplestore is more efficient if relations are specialized on the various classes. Partitions in various types of entities, namely Parameters and EntityData help to speed-up the search. Relations in the model can be qualified by attributes (the typical association class in UML). In the triplestore this requires an additional predicate. As an example, the relation "wasAttributedto" between an Entity and Agent has an extra predicate : ‘holdsRoleInTime‘ which allows for an Agent to play various roles w.r.t this entity: operator, author, provider, etc.

The number of triples is not limited and can be increased easily with the number of instances. The triple is a flat and additive representation and can be used to extend the model, for instance by adding new properties via new attributes to classes or to relations , or adding new relations between existing class instances , etc. Then tracing the evolution of provenance metadata in a project is available and at the same time, the compatibility to current IVOA Provenance data can co-exist.

A scaling test has been performed on Blazegraph implementing an excerpt of the CDS SIMBAD data base with a set of 20 relations. Successive increase of the number or of astronomical SIMBAD objects showed that Blazegraph has scaled properly for 10 000, 1million and 8,5 millions of objects. In order to speed-up the positional search for those objects a positional index based on HEALPIX coordinates representation has been successfully applied.

Other triplestore platforms will be tested to complete the tests.

Acknowledgments. This work has been sponsored by the DADI WP from the european ASTERICS project (Horizon 2020) and the CDS internship program.

References

Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. 2013, Prov-o: The prov ontology, W3C Recommendation. URL http://www.w3.org/TR/prov-o/

Dowler, P., Rixon, G., & Tody, D. 2010, Table Access Protocol Version 1.0, IVOA Recommendation 27 March 2010. 1110.0497