Provenance Tools for Astronomy

Michèle Sanguillon,[1] François Bonnarel,[2] Mireille Louys,[2,3] Markus Nullmeier,[4] Kristin Riebe,[5] and Mathieu Servillat[6]
[1]*Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS/IN2P3, France;* Michele.Sanguillon@umontpellier.fr

[2]*Centre de Données astronomiques de Strasbourg, Observatoire Astronomique de Strasbourg, Université de Strasbourg, CNRS, Strasbourg, France*

[3]*ICube Laboratory, Université de Strasbourg, CNRS, Strasbourg, France*

[4]*Zentrum für Astronomie der Universität Heidelberg, Astronomisches Rechen-Institut, Heidelberg, Germany*

[5]*Leibniz Institute for Astrophysics Potsdam, Germany*

[6]*Laboratoire Univers et Théories, Observatoire de Paris, PSL Research University, CNRS, 92190 Meudon, France*

Abstract. In the context of astronomy projects, scientists have been confronted with the problem of describing in a standardized way how their data have been produced.

As presented in a talk at last year's ADASS, the International Virtual Observatory Alliance (IVOA) is working on the definition of a Provenance Data Model, compatible with the W3C PROV model, which shall describe how provenance metadata can be modeled, stored and exchanged in astronomy.

In this poster, we present the current status of our developments of libraries and tools, mainly open source, which implement the IVOA Provenance Data Model in order to produce, serve, load and visualize provenance information. These implementations are also needed to validate and adjust the data model and the standard definitions for accessing provenance. The provenance tools developed and created for the W3C framework are reused and extended when possible to tackle the domain of astronomical data.

1. Introduction

The International Virtual Observatory Alliance[1] has developed several data models to foster interoperability between diverse astronomy projects. Even though a lot of objects (spectra, images, simulations, etc.) are already well described, some parts of the information about how datasets have been produced is still missing.

That is why the IVOA Data Model Working Group investigates how to model provenance information of a dataset, how this information can be stored and how it can be exchanged. In order to check the validity of the defined model, the group imple-

[1] http://www.ivoa.net/

mented the IVOA Provenance Data Model in four environments: Pollux, CTA, RAVE, and one at CDS.

Here, we present the tools developed to implement this model in these different contexts.

2. IVOA Provenance Data Model

The IVOA Provenance Data Model (Riebe et al. 2017) follows the W3C Provenance definition, i. e., that provenance is "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness".

The main core classes (*Entity*, *Activity*, *Agent*) and its relations (*wasGeneratedBy*, etc.) have the same name as in the W3C Provenance Data Model (Belhajjame et al. 2013). We add the *ActivityFlow* class and the *hadStep* relation in order to allow users to describe workflows of activities. We also add the possibility to separate the description of an activity or entity from the activity/entity itself.

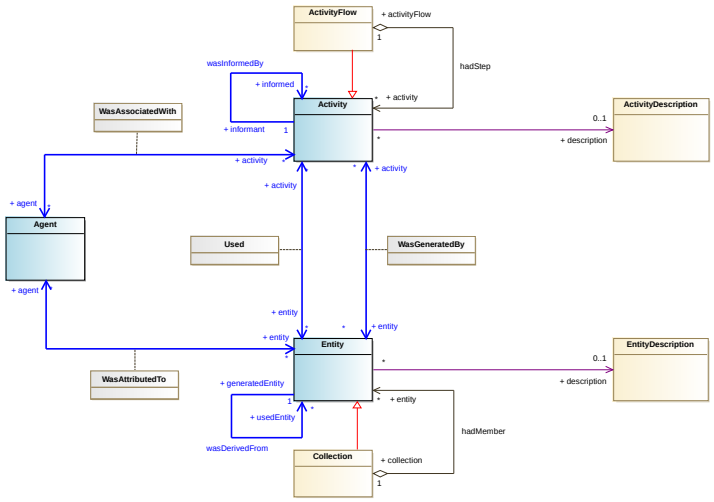

Figure 1. IVOA Provenance Data Model

3. voprov library

The voprov[2] package is an open source Python library derived from the prov[3] Python library (MIT license) developed by Trung Dong Huynh (University of Southampton).

The voprov package implements the serialization of the IVOA Provenance Data Model. As this model is very close to the W3C one, the voprov library uses the following facilites from prov: the PROV-N, PROV-JSON, and PROV-XML serialization

[2] `https://github.com/sanguillon/voprov/`

[3] `https://github.com/trungdong/prov/`

formats, as well as PDF, PNG, and SVG graphical representations. It adds these IVOA features: flows of activities (pipelines), which are composed of different activity steps, and serialization into the VOTable format.

This library is currently used in the context of the POLLUX database, which offers high resolution synthetic spectra computed using the best available models of the atmosphere and efficient spectral synthesis codes.

When a spectrum is integrated into the database, provenance information is retrieved and serialized in different formats and with different levels of detail. When a user or a program queries the Pollux database (via the SSA protocol of the Virtual Observatory), he is informed (via the DataLink protocol) of the existence of a service that allows him to retrieve provenance information in a given format and for a given detail level. This functionality has been implemented in the CASSIS spectrum visualization tool.

4. Django package

The django-prov_vo package[4] is an open source Python package that can be reused in Django web applications for serving provenance information. The data model classes are directly mapped to tables in a relational database. The package provides different interfaces to extract provenance: a REST interface to retrieve lists of entities, activities and agents, and a ProvDAL interface, which is defined in the current IVOA Provenance Working Draft. The ProvDAL interface takes the identifier of an entity, activity or an agent as a parameter and then returns the available provenance information in one of the serialization formats (currently PROV-N and PROV-JSON). A few visualization techniques for the retrieved provenance graph are also included.

This `django-prov_vo` package was developed for a provenance service of the RAVE[5] project. Within the RAVE (RAdial Velocity Experiment) survey, spectra of about half a million stars from the southern hemisphere were observed and stellar properties determined.

5. Prototype PostgreSQL database at CDS

We implemented the IVOA Provenance DM in a test Postgres database at CDS. The database handles a small collection of image datasets, such as Schmidt plates, monoband and color composed images or HiPS representations of pixel data. From the IVOA Provenance Datamodel specification we designed a database schema and implemented the various related tables recommended in the data model as Postgres tables.

A small set of plates, with their digitization, cutout extractions, RGB color composition, and HiPS generation activities, is used to populate the database. Various scenarios for querying and displaying their provenance information have been tested in SQL. For query responses, PROV-N, PROV-JSON, and PROV-VOTable formats are provided. A simple Python API allowing users to select the main types of requests and to display the responses via W3C Prov library has been designed. It allows users

[4] `https://github.com/kristinriebe/django-prov_vo`

[5] `https://www.rave-survey.org/`

querying for various combinations of provenance relationships in the database and to visualize the provenance graph in a user friendly representation.

This provides experience with the DM implementation and clues to build up a TAP SCHEMA representation for ProvTAP services, a preliminary version of which has been developed.

6. UWS Server at Observatoire de Paris

In the context of the Cherenkov Telescope Array[6] (CTA) project, a job control system based on the IVOA UWS pattern has been developed as an open source Python application: OPUS[7] (Observatoire de Paris UWS System). This system has been used to test the execution of CTA data analysis tools on a work cluster. It implements the ProvenanceDM concept of ActivityDescription files and provides the provenance information for each executed job in PROV-JSON and PROV-XML serializations.

The CTA is the next generation ground-based very high energy gamma-ray instrument. Contrary to previous Cherenkov experiments, it will serve as an open observatory providing data to a wide astrophysics community, with the requirement to offer self-described data products to users that may be unaware of the Cherenkov astronomy specificities (see also Servillat et al. (2018)).

Acknowledgments. This work was partially funded by the Federal Ministry of Education and Research in Germany and by the ASTERICS project (`http://www.asterics2020.eu/`). Additional funding was provided by the INSU (Action Spécifique Observatoire Virtuel, ASOV), the Grand-Sud-Ouest Data Centre, the Paris Astronomical Data Centre, and the Observatoire Astronomique de Strasbourg.

References

Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., & Tilmes, C. 2013, PROV-DM: The PROV data model, W3C Recommendation. URL `http://www.w3.org/TR/prov-dm/`

Riebe, K., Servillat, M., Bonnarel, F., Louys, M., Nullmeier, M., Rothmaier, F., Sanguillon, M., & the IVOA Data Model Working Group 2017, IVOA provenance data model, `http://www.ivoa.net/documents/ProvenanceDM/`

Servillat, M., Boisson, C., Lefaucheur, J., Kosack, K., Sanguillon, M., Louys, M., & Bonnarel, F. 2018, in ADASS XXVII, edited by TBD (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD

[6] `https://www.cta-observatory.org/`

[7] `https://github.com/mservillat/OPUS`