A Provenance Data Model for Astronomy

Kristin Riebe,[1] Francois Bonnarel,[2] Mireille Louys,[2] Florian Rothmaier,[3] Michèle Sanguillon,[4] and Mathieu Servillat,[5]

[1]*Leibniz-Institute for Astrophysics Potsdam, Potsdam, Germany;* kriebe@aip.de

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

[3]*LSW, Zentrum für Astronomie Heidelberg, Heidelberg, Germany*

[4]*University of Montpellier, LUPM, Montpellier, France*

[5]*LUTH, Observatoire de Paris, PSL Research University, CNRS, France*

Abstract. In Astronomy as well as in other sciences it is of crucial importance to have information about the origin and history of data, i.e. its provenance information. The IVOA Provenance Data Model shall record this information in a consistent and interoperable way for astronomical data. This will enable scientists to use common tools for discovering data based on their provenance and help them to gain a better understanding of the data, its processing and to judge the data's quality and reliability.

1. Introduction

Provenance information for astronomical data products is metadata about its origin, i.e. how the data was produced, which other datasets played a role in its production and who was responsible for it. If such information is available for each (published) science-ready dataset, scientists can better evaluate if the data is suitable for his/her specific research goals. It is also possible to go further and provide provenance information for each plot published in a paper, enabling scientists to recover the origin of each individual point of e.g. a light curve or a spectral energy distribution.

By defining a Provenance Data Model, the IVOA is therefore making an effort to enhance reproducibility and understanding of data products as well as allowing interoperability of provenance information. The following section will outline the main important ideas and concepts of the current working draft (see Riebe et al. 2016).

2. Provenance Data Model

The IVOA Provenance Data Model for astronomical data has three core classes. The names of these classes are the same as the corresponding classes from the W3C provenance data model (Belhajjame et al. 2013) for maximizing compatibility with the world outside of astronomy and taking benefits of existing tools provided around the W3C Provenance initiative (Moreau et al. 2011). These core classes are:

- *Activity*: a process, performing a certain task over a period of time
- *Entity*: a thing or data product, e.g. a table, image, spectrum, parameter
- *Agent*: a person or a project/organization responsible for an activity or entity

A workflow in astronomy typically can be described by a chain of activities and entities: a process (e.g. an observation) produces datasets (raw images), which are then input for another process (the pipeline) which produces a reduced dataset (science-ready data). Looking at the provenance of data, one would recover this process backwards: given a piece of data from a data release, a scientist asks for the process that generated this dataset, which input data it was using, which processes where generating these etc. Figure 1 depicts such a simple "backwards workflow".

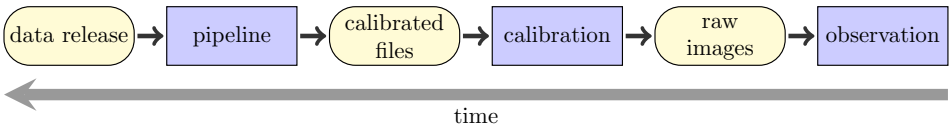

Figure 1. Schematic simplified example workflow in astronomy as it could be discovered using provenance information. Provenance discovery happens usually backwards in time. The (blue) boxes mark activities, the rounded (yellow) ones are entities.

For each of these steps different people and projects may be involved, for example the organization operating the telescope, the observer taking images, the software developer who wrote the pipeline or the project for which the observations were done (e.g. RAVE survey). Of course, in reality, many more such processes may play a role, and datasets of different origin may be combined in different ways to get certain results. But they all have in common that output data of one process may be used as input data for another process.

Just identifying the involved entities and activities is not enough. We also need to define the relations between them. Since we realized that we need the same relations as in the W3C provenance model, we reuse again the W3C terminology:

- *WasGeneratedBy*: an entity was generated by an activity
- *Used*: an activity used one or many (input) entities
- *WasAssociatedWith*: an agent was associated with an activity
  (e.g. someone performed an observation, wrote or applied some software)
- *WasAttributedTo*: an agent was attributed to a dataset
  (e.g. the RAVE project gets attribution for the RAVE Data Release 4)

These core classes are already sufficient to trace basic provenance information. But if processes are being repeatedly applied to different datasets (as is done quite often in astronomy), then it is more convenient to store descriptions for them in an extra *description* class and make a reference from each corresponding activity to this common description. The same applies to entities: if many entities with the same description have to be recorded, then one can define a common entity description. All these classes and relations are put together in Figure 2.

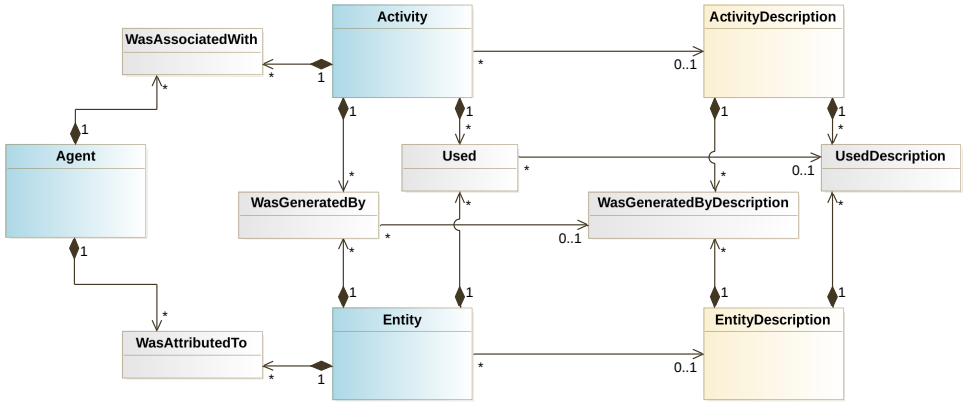

Figure 2. The main UML classes of the Provenance Data Model and relations between them. Core classes that also correspond to classes in the provenance model of W3C are colored blue (Activity, Entity, Agent), relation-classes are colored in grey. The description classes on the right hand side are optional.

3. Links to other data models

IVOA has other data models for specific areas, e.g. ObsCore (Louys et al. 2011) for observational data or the Simulation Data Model (SimDM, Lemson et al. 2012), which covers the description and even the provenance of simulations. The entities and their descriptions in the Provenance Data Model are tightly linked to the ObsDataSet-class in the ObsCore data model, as well as to InputDataset and OutputDataSet in SimDM. The concept of normalized description classes is also known in SimDM, where the actual execution of a simulation is called "experiment" (a type of Activity) and its general description is called a "protocol" (ActivityDescription). Provenance information is already partially covered in other data models and we currently develop an appropriate mapping for linking the Provenance Data Model to existing IVOA data model elements.

Since we use the same class names as the W3C model where possible, one can also directly generate W3C compatible serializations of the model, so that existing tools for viewing and storing W3C provenance can be reused.

4. Use cases

Several use cases have been identified and are explained in more detail in the IVOA Provenance Data Model Working Draft (Riebe et al. 2016); the Cherenkov Telescope Array (CTA) use case is also explained in the poster Servillat, Mathieu (2017) of this conference. Provenance plays a major role for CTA, because after a particle shower was observed, simulations are used to then reconstruct the original signal that caused the particle shower in the atmosphere. Understanding which simulations have been used is crucial for scientists in order to correctly use the final data products for their science cases.

The POLLUX database contains high resolution synthetic spectra and currently stores provenance information in a non-standardized way. Implementing this Provenance Data Model will make it easier to exchange and visualize provenance descriptions, and will allow users to select datasets based on provenance criteria.

These are just two examples where provenance is needed and the current model is already being implemented. Such implementation experiences help to check our model for usefulness and to discover where further improvements are needed.

5. Summary

In this document we briefly outlined the current status of the IVOA Provenance Data Model, which is still in development. It is closely related to the W3C Provenance Data Model, especially concerning the core elements, with additional description classes for separating reusable descriptions from the actual data entities or executions of processes. Released versions and latest working drafts of the IVOA Provenance specification are available on the IVOA website in the documents section[1].

Acknowledgements. This work was developed with support from the German Astrophysical Virtual Observatory, funded by BMBF Bewilligungsnummer 05A08VHA and 05A14BAD, and with support from the ASTERICS Project, funded by the European Commission (project 653477).

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/

Lemson, G., Wozniak, H., Bourges, L., Cervino, M., Gheller, C., Gray, N., LePetit, F., Louys, M., Ooghe, B., & Wagner, R. 2012, Simulation Data Model Version 1.0, IVOA Recommendation 03 May 2012. 1402.4744

Louys, M., Bonnarel, F., Schade, D., Dowler, P., Micol, A., Durand, D., Tody, D., Michel, L., Salgado, J., Chilingarian, I., Rino, B., de Dios Santander, J., & Skoda, P. 2011, Observation data model core components and its implementation in the Table Access Protocol, version 1.0, IVOA Recommendation. URL http://www.ivoa.net/documents/ObsCore/20111028/REC-ObsCore-v1.0-20111028.pdf

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., & den Bussche, J. V. 2011, Future Generation Computer Systems, 27, 743 . URL http://www.sciencedirect.com/science/article/pii/S0167739X10001275

Riebe, K., Servillat, M., Bonnarel, F., Louys, M., Rothmaier, F., Sanguillon, M., & IVOA Data Model Working Group 2016, IVOA Provenance Data Model, IVOA Working Draft. URL http://www.ivoa.net/documents/ProvenanceDM/

Servillat, Mathieu 2017, in ADASS XXVI, edited by TBD (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD

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