Provenance for astrophysical data*

Anastasia Galkin[1][0000−0003−0131−7491] agalkin@aip.de, Kristin Riebe[1], Ole Streicher[1][0000−0001−7751−1843], Francois Bonnarel[2], Mireille Louys[2], Michèle Sanguillon[3][0000−0003−0196−6301], Mathieu Servillat[4][0000−0003−0196−6301], and Markus Nullmeier[5]

[1] Leibniz-Institute for Astrophysics Potsdam (AIP), Germany
[2] Universitè de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, France
[3] LUPM, CNRS, Université de Montpellier
[4] Laboratoire Univers et Thèories, Observatoire de Paris, PSL Research University, CNRS, France
[5] ARI, Zentrum fuer Astronomie Heidelberg, Heidelberg, Germany

Abstract. In the context of astronomy projects, provenance information is important to enable scientists to trace back the origin of a dataset. It is used to learn about the people and organizations involved in a project and assess the quality of the dataset as well as the usefulness of the dataset their scientific work. As part of the data model group in the International Virtual Observatory Alliance (IVOA) we are working on the definition of a provenance data model for astronomy which shall describe how provenance metadata can be modeled, stored and exchanged. The data model is being implemented for different projects and use cases.

Keywords: astronomy · astrophysics · escience · data management · provenance · IVOA.

1 Introduction

The Virtual Observatory (VO) is the vision that astronomical datasets and other resources should work as a seamless whole. The IVOA [4] is an organisation that debates and agrees on the technical standards that are needed to make the VO possible.
The goal of the IVOA Data modeling group is to develop a provenance data model which will not only store provenance information but also to find ways to let the astronomical community explore provenance in a interoperable way, linking the provenance information to already existing VO data models and infrastructures.

* This project is partially funded by BMBF 05A14BAD and 05AI7BA2S. Additional funding is provided by ASTERICS (http://www.asterics2020.eu/), a project supported by the European Commission Framework Programme Horizon 2020 Research and Innovation action under grant agreement n.653477. Further funding was provided by the German Virtual Observatory (GAVO), the French Virtual Observatory (ASOV OV-France), and Paris Astronomical Data Centre (PADC).

2 Use cases for provenace in astronomy

For an astronomical data set, provenance can answer questions such as: Which processing steps have been done already? Who was involved in the project? Who can I ask about this data? Is the dataset suited for my research? Which datasets were produced with the same pipeline version? "Forward tracking" is useful to follow the usage within the given domain. Structured provenance metadata helps to find possible error sources such as the version of processing software, telescope configuration, parameter settings.

2.1 Cherenkov Telescope Array

The Cherenkov Telescope Array (CTA)[3] is the next generation ground-based very high energy gamma-ray instrument. Contrary to previous Cherenkov instruments, it will serve as an open observatory providing data to a wide astrophysics community, with the requirement to propose self-described data products to users that may be unaware of the Cherenkov astronomy specificities. Provenance is used to organize the data reprocessing workflow of the pipeline.

2.2 Spectroscopic surveys

In large spectroscopic surveys (e.g. 4MOST [1]), sections of the sky are scanned to retrieve characteristics of the electromagnetic radiation emmitted by cosmological objects, such as stars, black holes or galaxies. The provenance model can help identify adjacent objects on the CCD (an electronic light sensor) for a given 1D spectrum to identify sources of crosstalk.

2.3 APPLAUSE database - scanning historical photoplates

The APPLAUSE archives [2] host digitized copies of photographic plates from the German astronomical observatories. These items are of particular interest for the study of long-term variability of many types of stars. The provenance use-case here encompasses physical objects such as photographic plates, the scanners and the log book as well as the software processing steps, parameters and the digital outcome of the project - digitized images and identified objects such as stars. The data release 3 is planned to be published in July 2018, provenance metadata are being constructed and will be added in a later addition.

2.4 MUSE Data Reduction Pipeline

The Multi Unit Spectroscopic Explorer (MUSE) [5] is an instrument installed at the Very Large Telescope (VLT) of the European Southern Observatory. The raw data are recorded separately and then transformed into a fully calibrated, science-ready data cube using the MUSE data reduction pipeline [13]. All information is stored in a specific object oriented database [12]. First attempts to describe the provenance information using the W3C model for one final data cube result in metadata containing about 2700 file entities and 270 activities (recipe runs).

2.5 RAVE survey

RAVE (Radial Velocity Experiment) [6] is one of the largest spectroscopic surveys of Milky Way stars. The final data products are data release tables with properties for half a million stars. These properties are derived from the original raw spectra which are observed by a number of fibres attachedf to the telescope and were processed in numerous processing steps. If the provenance information contains all the details and intermediate steps, tracking back the provenance for each stellar property through to the original fibre-spectrum, the amount of information becomes overwhelming.

3 Special requirements in modelling provenance in astronomy

The IVOA provenance data model follows closely the W3C provenance model [7], utilizing entities, activities and agents, and the relevant relations between them. The provenance information use in astronomy has however some specific challenges:
First, the astronomical provenance records are highly complex. A coarse or a detailed view of a provenance model is needed depending on the task where the provenance is used.
Second, many tasks in astronomy are repetitive, e.g. several observations can be performed with the same telescope and instrument, or many simulations are performed using the same code and computing environment, but with slightly varying code parameters. This is normalized by using a special class that abstractly describes the activity. The complex data processing also may require to structure the workflow by combining several activities into one.
Last, activities highly rely on parameters and parameter sets. Parameters have a value and might or might not have a history as well. Thus, parameters could be modeled as entities.

4 Integration into the IVOA ecosystem

IVOA has built up a well functioning and widely used ecosystem of interoperable services and tools such as Tool for OPerations on Catalogues And Tables (TOPCAT). One of the main concepts in IVOA is the Table Access Protocol (TAP) [8]. Within the TAP protocol the access is provided for both the database and the table metadata as well as for actual table data. TAP also includes support for synchronous and asynchronous queries as well as support for multiple query languages, mainly the Astronomical Data Query Language (ADQL).
The ProvTAP accesses provenance information accordingly to the TAP standard. The output format is VOTable, the VO standard table output format.
ProvSAP (for Simple Access Protocol) allows the client to request information in a REST framework way with W3C output formats such as PROV-JSON, PROV-XML and PROV-N.

Provenance information can also be directly stored in data files such as images files (FITS) or in VOTables. The standard for it is currently discussed by the IVOA modeling group.

5 Summary

In this document we briefly outlined the development of the IVOA provenance model for the astronomical scientific field. The current IVOA Provenance Data Model is still in development and some core concepts are in discussion now. We welcome and encourage the input of W3C provenance experts to complete the model within the IVOA ecosystem.
For further reading please look at various proceedings, documents and notes, e.g.: [10], [11].
The latest official version of the working draft can be found at [9]. The released versions will be published at the IVOA website [4] in the documents section.

References

1. The 4-metre multi-object spectrograph telescope, https://www.4most.eu/
2. Applause - archives of photographic plates, https://plate-archive.org
3. Cherenkov telescope array, https://www.cta-observatory.org/
4. International virtual observatory alliance (ivoa), http://ivoa.net
5. Muse science - the multi unit spectroscopic explorer, https://muse-vlt.eu/science
6. Rave – the radial velocity experiment, https://www.rave-survey.org/project/
7. 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.: PROV-DM: The prov data model. W3C Recommendation (Apr 2013), http://www.w3.org/TR/prov-dm/
8. Dowler, P., Rixon, G., Tody, D., Demleitner, M.: Table access protocol - version 1.1. http://www.ivoa.net/documents/TAP/ (2018)
9. Riebe, K., Servillat, M., Bonnarel, F., Galkin, A., Louys, M., Nullmeier, M., Rothmaier, F., Sanguillon, M., Streicher, O., the IVOA Data Model Working Group: IVOA provenance data model. http://www.ivoa.net/documents/ProvenanceDM/ (2017)
10. Riebe, K., Servillat, M., Bonnarel, F., Louys, M., Sanguillon, M., the IVOA Data Model Working Group: Provenance implementation note. http://volute.g-vo.org/svn/trunk/projects/dm/provenance/implementation-note/ (2017)
11. Servillat, M., Boisson, C., Lefaucheur, J., Kosack, K., Sanguillon, M., Louys, M., Bonnarel, F.: Provenance as a requirement for large-scale complex astronomical instruments. In: ADASS XXVII. ASP Conf. Ser., ASP, San Francisco (2018)
12. Vriend, W.J.: Porting Big Data technology across domains. WISE for MUSE. In: Science Operations 2015: Science Data Management - An ESO/ESA Workshop, held 24-27 November, 2015 at ESO Garching. p. 1 (Dec 2015). https://doi.org/10.5281/zenodo.34624
13. Weilbacher, P.M., Streicher, O., Urrutia, T., Pécontal-Rousset, A., Jarno, A., Bacon, R.: The MUSE Data Reduction Pipeline: Status after Preliminary Acceptance Europe. In: Manset, N., Forshay, P. (eds.) Astronomical Data Analysis Software and Systems XXIII. Astronomical Society of the Pacific Conference Series, vol. 485, p. 451 (May 2014)