PROCEEDINGS OF SCIENCE

Archiving data from a software telescope

Catherine Boisson[1], **Mathieu Servillat**
LUTh, PADC, Observatoire de Paris/PSL/CNRS/Université Paris-Diderot
Meudon, France
E-mail: catherine.boisson@obspm.fr, mathieu.servillat@obspm.fr

Karl Kosack
IRFU, CEA, Université Paris-Saclay
91191 Gif-sur-Yvette, France
E-mail: karl.kosack@cea.fr

Mireille Louys
ICube Laboratory, Université de Strasbourg, CNRS
67000 Strasbourg, France
E-mail: mireille.louys@unistra.fr

François Bonnarel
CDS, Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550
11 rue de l'Université, 67000 Strasbourg, France
E-mail: francois.bonnarel@astro.unistra.fr

Michèle Sanguillon
Laboratoire Univers et Particules de Montpellier, Université de Montpellier
CNRS/IN2P3, France
E-mail: michele.sanguillon@umontpellier.fr

The landscape of ground-based gamma-ray astronomy is drastically changing with the perspective of the Cherenkov Telescope Array (CTA). For the first time in this energy domain, CTA will be operated as an observatory open to the astronomy community and produce data that will be publicly released to a large community of scientists. In the context of Cherenkov astronomy, the data processing stages imply both assumptions and comparison to dedicated simulations. As a consequence, "Provenance" information is crucial to the end user in order to interpret the high level data products and there are thus strong requirements to ensure data quality, reliability and trustworthiness. Among those requirements, traceability and reproducibility of the data products can be answered by structuring and storing the provenance information for each data product. We are partners in ASTERICS DADI and developed several pieces of software to enable the tracking of provenance information for the large-scale complex astronomical observatory CTA and a web-based data diffusion prototype, in close relation with the International Virtual Observatory Alliance (IVOA).

The New Era of Multi-Messenger Astrophysics - Asterics2019
25 – 29 March, 2019
Groningen, The Netherlands

[1]Speaker

PoS(Asterics2019)015

1. Introduction

Scientists are confronted with the problem of describing in a standardized way how their data have been produced, a crucial problem in the context of astronomy projects where large-scale complex astronomical instruments and observation data bases are now being build. The development of large observatories, as it is the case for example for the Cherenkov Telescope Array[2] (CTA), is from consortia of specialized physicists. The path of the data production from acquisition to dissemination, through e.g. data centres, archives and web portals, can be extremely obscure to the end user.

Contrary to previous ground-based 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. In the context of Cherenkov astronomy, the data processing stages imply both assumptions and comparison to dedicated simulations. Publishable data are thus not direct observables but rely on the many assumptions during the full processing chain. To assess the *usefulness* and the *quality* of the data for their own scientific work, end users need a flowchart explaining the large number of steps and complexity involved in the data preparation. This can be done by collecting provenance information at each step of the data preparation.

For that we follow and actively participate to the design of the IVOA Provenance data model[3] (Servillat et al. 2019) to store provenance information during data production, and implemented solutions to collect provenance information during the CTA data processing and the execution of jobs on a work cluster.

2. Provenance information during the CTA data production

The production of scientific data from CTA uses a complex and specific Pipeline[4], accessing different resources and calibration products, and using complex algorithms. Ensure that the data processing is traceable and reproducible is a requirement for the CTA Pipeline design. The storage of provenance information at each step of the data processing is a key feature of the Pipeline, and was introduced early in the data model design.

The IVOA Provenance Data Model 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” (Belhajjame et al. 2013). However the W3C core model being too generic, new classes such as ActivityDescription, EntityDescription, Parameter/ParameterDescription have been added, along with relations, types and roles in order to allow users to describe in details the processes and method used behind each activity. We also add the possibility to separate the description of an activity or entity from the activity/entity itself, and attach ActivityConfiguration information (what parameters were used so that the activity occurred in the desired conditions).

[2] http://www.cta-observatory.org/
[3] http://www.ivoa.net/documents/ProvenanceDM/
[4] https://github.com/cta-observatory/ctapipe

The granularity is chosen in order to allow full reproducibility of the process, and the relevant information attached to datasets allow to estimate their pertinence for a scientific use.

All relevant metadata from the CTA data model are included (Servillat et al. 2017). The Provenance information is collected at each step of the data processing. Unique identifiers are used for entities, activities and agents; a list of all used and generated entities during the execution of an activity are kept. A Provenance Python class has been developed for the CTA Pipeline. This class is automatically loaded when a task is executed and provenance information is automatically recorded : when the task is started, when it ends, when an input entity (file, database access) is touched and when an output entity is created. This makes the collection of provenance information mostly hidden to the user, but also to the developers. The resulting dictionary at the end of the task could be combined with a description of the task to generate an IVOA Provenance compatible file, adding in particular links to persons responsible for the task.

The Provenance class serves different goals, first the tracking of the history of a data product to inform the end user about its origin and quality, but also the possibility to check the integrity of the Pipeline an locate sources of errors by searching structured provenance information.

3. Store and expose the provenance information

To allow recovery of the provenance, a light job control system that stores provenance information based on the IVOA UWS pattern and Provenance data model, OPUS[5] (Observatoire de Paris UWS System), was developed as an open source Python application (Servillat et al., 2018). This system was used to test the execution of CTA data analysis tools on a work cluster. It implements the Provenance DM concept of ActivityDescription files and provides the provenance information for each executed job in different serializations. This information is attached to the job results and can be visualized as a graph. OPUS is a reference implementation for the IVOA Provenance data model which provides a test platform for the management of provenance information.

4. Conclusion

We successfully followed the Provenance data model, currently discussed at the IVOA, and implemented solutions to collect provenance information during the CTA data processing and the execution of jobs on a work cluster. The active participation of CTA in the definition of the IVOA model for Provenance, and the inclusion of the model in CTA data management system to define Provenance Configuration show that the Provenance DM is now mature.

The various profiles the IVOA DM can offer :

– Workflow : Activity focused

– Data flow / archive : Dataset focused

– Credits/responsibility views

All those are recoverable in different serializations.

Bring your Use Cases !

[5] https://github.com/mservillat/OPUS

References

[1] K. Belhajjame et al.. 2013, PROV-DM: The prov data model, W3C Recommendation. http://www.w3.org/TR/

[2] M. Servillat et al. 2017, Structuring metadata for the Cherenkov Telescope Array, ADASS XXVI proceedings (Trieste 2016), ASP Conf. Ser, arXiv:1706.06512

[3] M. Servillat et al., 2018, "Provenance as a requirement for large-scale complex astronomical instruments", ADASS XXVII proceedings (Santiago 2017), ASP Conf. Ser, arXiv:1806.00447

[4] M. Servillat et al., 2019, IVOA provenance data model, Working Draft, http://www.ivoa.net/documents/ProvenanceDM/

Acknowledgment

The authors acknowledge support from ASTERICS, a project funded by the European Commission under the Horizon2020 programme (id 653477), in the framework of ASTERICS Work Package 4 *Data Access, Discovery and Interoperability*. Additional funding was provided by the INSU (Action Spécifique Observatoire Virtuel, ASOV), the Action Fédératrice CTA at the Observatoire de Paris and the Paris Astronomical Data Centre (PADC).