Provenance as a requirement for large-scale complex astronomical instruments

Mathieu Servillat,[1] Catherine Boisson,[1] Julien Lefaucheur,[1] Karl Kosack,[2] Michèle Sanguillon,[3] Mireille Louys,[4,5] and François Bonnarel[4]

[1]*Laboratoire Univers et Théories, Observatoire de Paris, PSL Research University, CNRS, 92190 Meudon, France;* mathieu.servillat@obspm.fr

[2]*CEA Saclay, DSM/IRFU/SAp, Bat 709, F-91191 Gif-Sur-Yvette, France*

[3]*Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS/IN2P3, France*

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

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

Abstract.

We developed several pieces of software to enable the tracking of provenance information for the large-scale complex astronomical observatory CTA, the Cherenkov Telescope Array. Such major facilities produce data that will be publicly released to a large community of scientists. There are thus strong requirements to ensure data quality, reliability and trustworthiness. Among those requirements, traceability and reproducibility of the data products have to be included in the development of large projects. Those requirements can be answered by structuring and storing the provenance information for each data product.

We 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.

1. Introduction

State of the art observations are now performed by large-scale complex astronomical instruments. A consortium of specialists is generally responsible for the development and the operation of large observatories, as it is the case for example for the Cherenkov Telescope Array[1] (CTA). The path of the data production from acquisition to dissemination, through e.g. data centers, archives and web portals, can be extremely obscure to the end user. This complexity is illustrated in Figure 1.

In order 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

[1]http://www.cta-observatory.org/

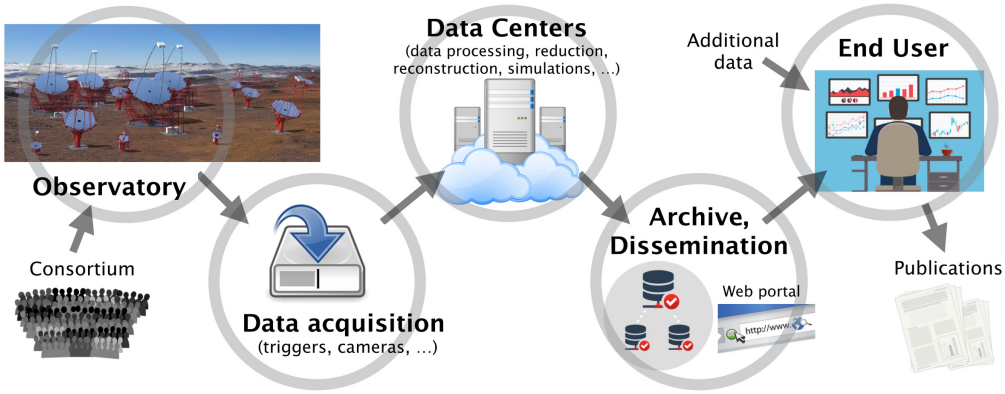

Figure 1. Data production for large-scale complex astronomical instruments such as CTA.

at each step of the data preparation. We followed the IVOA Provenance data model (Riebe et al. 2017; Sanguillon et al. 2018) to develop solutions for CTA.

Provenance is information about entities, activities, and people involved in producing a piece of data(Belhajjame et al. 2013). It helps to trace back the data lineage through the production pipeline, and learn about the methods used and the people or organizations involved in the project.

2. How to collect provenance information during CTA data production?

The production of scientific data from CTA will use a complex and specific Pipeline, accessing different resources and calibration products, and using complex algorithms. A key feature of the Pipeline that was included early in the development is the storage of provenance information at each step of the data processing.

In order to enable the recording of Provenance information, the development followed those steps, where it was important to include the notions of Provenance early in the data model design:

- Include the relevant metadata in the CTA data model (Servillat et al. 2017)
- Follow the IVOA Provenance data model for the generated data (Riebe 2017)
- Collect provenance information at each step of the data processing:
  - Use unique identifiers for entities, activities and agents
  - Describe each task executed in the Pipeline
  - Keep a list of all used and generated entities during the execution of an activity

A Provenance Python class has been developed for the CTA Pipeline framework `ctapipe`[2]. This class is loaded automatically when a task is executed and provenance

[2] `https://github.com/cta-observatory/ctapipe`

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. How to store and expose the provenance information in a standard format?

We developed a job control system that stores provenance information following the IVOA UWS pattern and Provenance data model. OPUS[3] (Observatoire de Paris UWS System) is a light job control system developed as an open source Python application.

The following features have been implemented:

- Edit and fill Activity Descriptions, following the proposed IVOA standard serialization (Riebe et al. 2017)
- Run jobs asynchronously on a work cluster. OPUS connects with the workload manager used at the Observatoire de Paris (SLURM – Simple Linux Utility for Resource Management), but it can also run jobs on the local computer/server.
- Present the list of jobs attached to a user per available job.
- Present a status page for each job with input and results.
- Generate and return Provenance files after job completion, that are attached to the job as results.

This system has been used to test the execution of CTA data analysis tools on a work cluster, as it can be seen in Figure 2. Such a service can be included in a data access web portal as it is currently tested in the CTA Data Distiller prototype[4].

4. Conclusion

We developed tools that implement the IVOA Provenance proposed standard in the context of a large-scale complex astronomical observatory, with the aim to provide generic tools that can be used for other projects.

---

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

[4] `https://voparis-cta-test.obspm.fr`

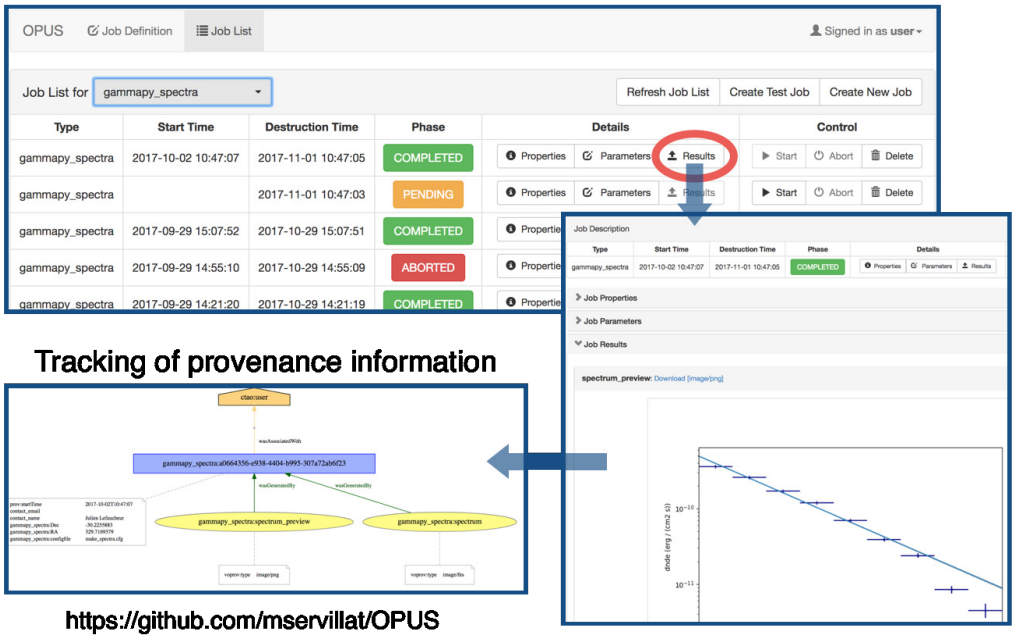

Figure 2. Screenshots of OPUS for a CTA related job. The joblist for the user and the job *gammapy_spectra* is shown at the top, then the result page with preview of the spectra is shown pn the right, and the provenance tree for this job is attached in PROV format on the left.

Acknowledgments. This work was partially funded 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; 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).

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

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/`

Sanguillon, M., Bonnarel, F., Louys, M., Nullmeier, M., Riebe, K., & Servillat, M. 2018, in ADASS XXVII, edited by TBD (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD

Servillat, M., Boisson, C., Lefaucheur, J., Brégeon, J., Sanguillon, M., Contreras, J.-L., & for the CTA Consortium 2017, in ADASS XXVI, edited by TBD (San Francisco: ASP), vol. TBD of ASP Conf. Ser., TBD. `1706.06512`