Practical Provenance in Astronomy

Mathieu Servillat, François Bonnarel, Mireille Louys, Michèle Sanguillon

Practical Provenance in Astronomy

Mathieu Servillat,[1] François Bonnarel,[2] Mireille Louys,[2,3] and Michèle Sanguillon[4]

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

[2]*Centre de Données astronomiques de Strasbourg, Observatoire Astronomique de Strasbourg, Université de Strasbourg, CNRS-UMR 7550, Strasbourg, France*

[3]*ICube Laboratory, Université de Strasbourg, CNRS-UMR 7357, Strasbourg, France*

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

Abstract.
Recently the International Virtual Observatory Alliance (IVOA) released a standard to structure provenance metadata, and several implementations are in development in order to capture, store, access and visualize the provenance of astronomy data products. This BoF will be focused on practical needs for provenance in astronomy. A growing number of projects express the requirement to propose FAIR data (Findable, Accessible, Interoperable and Reusable) and thus manage provenance information to ensure the quality, reliability and trustworthiness of this data. The concepts are in place, but now, applied specifications and practical tools are needed to answer concrete use cases. During this session we discussed which strategies are considered by projects (observatories or data providers) to capture provenance in their context and how a end-user might query the provenance information to enhance her/his data selection and retrieval. The objective was to identify the development of tools and formats now needed to make provenance more practical needed to increase provenance take-up in the astronomical domain.

1. The path to Open Science

Open science is the movement to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of an inquiring society, amateur or professional.

In this context, the European Open Science Cloud (EOSC) is an environment for hosting and processing research data to support science. The aim is to make research data interoperable and machine actionable following the FAIR principles[1] (Wilkinson

---

[1] `https://www.go-fair.org/fair-principles`

et al. 2016), by federating existing research data infrastructures in Europe and realising a web of FAIR data and related services for science, .

Within EOSC, the ESCAPE project[2] (European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures) brings together the astronomy, astroparticle and particle physics communities and puts together a cluster with ESFRI projects with aligned challenges of data-driven research. Keeping and exposing provenance information is one of those challenges.

A workshop[3] dedicated to the management of provenance took place in September 2020 within ESCAPE, in line with this ADASS session on practical provenance. The discussions led to a clarification of the requirement for structured provenance (§3), a more precise terminology (§4), and the identification of key concepts associated with the management of provenance information (§5).

2.   The IVOA Provenance Data Model

The IVOA has released a recommendation for a Provenance Data Model (Servillat et al. 2020b). The core model and definition of provenance is as defined by the W3C (Belhajjame et al. 2013): information about entities, activities, and people (agents) involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Provenance is related by definition to the origin of a product (where does it come from?), but also the path followed to generate this product (what has been done?). Provenance is thus seen as a chain of activities and entities, used and generated. With the core data model, the basic objectives are achieved: use of unique identifiers, traceability of the operations, connection with contacts for further information, citation or acknowledgement. By following the full IVOA data model, more advanced questions are answered: What happened during each activity? How was the activity tuned to be executed properly? What kind of content is in the entities?

The data model is a base for the development of tools and services, see e.g.: Servillat et al. (2021); Sanguillon et al. (2021); Landais et al. (2021); Servillat et al. (2020a); Sanguillon et al. (2020).

3.   Requirement for structured provenance

We often realize too late that there are missing elements or links in the provenance. The capture of the provenance should thus be structured, as detailed as possible, and as naive as possible (simply record what happens).

There are clear advantages to retain this information as structured, machine-readable data, in particular in the context of Open Science:

- Explicitly required through the **FAIR principles** that indicate to include **rich metadata**, following standard data model, protocols and formats, with **detailed provenance**.

---

[2]`https://projectescape.eu`

[3]`https://indico.in2p3.fr/event/21913/page/2641-summary`

- **Quality / Reliability / Trustworthiness** of the products: the simple fact of being able to show its provenance is sufficient to give more value to a product, and if the provenance information is detailed, the value will be higher.
- **Reproducibility requirement** in many projects: provenance details are essential to be able to rerun each activity (maybe testing and improving each step); Having this information, it may not be necessary to keep every intermediate file that is easily reproducible (hence a possible gain on storage space and costs).
- **Debugging**: with detailed provenance, it is not necessary to restart from scratch, as one can locate in the provenance graph the faulty parts or the products to be discarded, and reprocess only from the identified failing steps.

4. Some terminology

The word "provenance" is used to refer to different aspects depending on the persons or the goals involved. Provenance may be used for internal data management, or to improve the scientific exploitation of a data product. It may be stored inside the data file or separately (external file or database). We propose here a base for the definition of provenance categories.

- **full provenance**: graph/tree/chain of activities and entities up to the raw data. This information is not hosted by the entities themselves, but stored on an external server, or as separate files.
- **minimum provenance**: information attached to an entity as a list of keywords that gives some context and info on last activity (general process/workflow, software versions, contacts...), maybe including the list of used entities, so that a full provenance may be reconstructed from minimum provenance, but such information is not always preserved.
- **end-user/specific "provenance"**: information attached to an entity, as a list of keywords or data that provides key information to use/analyse the entity (e.g. for CTA: event class, event type, telescope configuration, sky conditions, reconstruction method,...). This information may be extracted from the full provenance, or inversely, used to enrich a reconstructed provenance.

5. Applying the model

Different concepts were identified concerning provenance in various use cases. We distinguish **on-top** provenance handling, where data products/collections already exist and is used to reconstruct a provenance graph, and **inside** provenance processing, where the provenance information is saved during the execution of the processing activities.

The use of **unique identifiers** is recommended and implies to think about provenance capture at the conception phase of a project. One also has to evaluate in advance the necessary and sufficient **granularity** (what steps? what objects?) and **level of details** (inclusion of descriptions and configuration details?).

The challenges in provenance management encompass the **capture**, **storage**, **access**, and **visualization** of the provenance information.

6.   Discussion topics

The discussion was organised around several topics, following the priorities of the participants as reported in a preliminary questionnaire. Those topics will be discussed during forthcoming ESCAPE events, in collaboration with IVOA working groups:

1. Defining the content of a minimum provenance, with a list of keywords related to the last activity and context.

2. Serializing provenance, both in a human- and machine-readable way.

3. Provenance and workflows, with workflow information simply attached to provenance (as used entities), or mapped to the model.

4. Evolving from provenance "on-top" to provenance "inside", e.g. with generic tools to capture provenance at the execution.

5. Provenance storage in a database and related interfaces.

6. Provenance exploration and visualization, based on access protocols (ProvTAP, see Bonnarel et al. 2019, or ProvSAP, see Servillat et al. 2021.), and tools such as the `voprov` Python package.

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/`

Bonnarel, F., Louys, M., Mantelet, G., Nullmeier, M., Servillat, M., Riebe, K., & Sanguillon, M. 2019, in ADASS XXVIII, edited by P. J. Teuben, M. W. Pound, B. A. Thomas, & E. M. Warner, vol. 523 of ASP Conf. Ser., 313

Landais, G., Servillat, M., Bonnarel, F., Louys, M., Sanguillon, M., & Michel, L. 2021, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici, vol. TBD of ASP Conf. Ser., 999 TBD

Sanguillon, M., Arrabito, L., Boisson, C., Bregeon, J., Kosack, K., & Servillat, M. 2021, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici, vol. TBD of ASP Conf. Ser., 999 TBD

Sanguillon, M., Bonnarel, F., Louys, M., Nullmeier, M., Riebe, K., & Servillat, M. 2020, in ADASS XXVII, edited by P. Ballester, J. Ibsen, M. Solar, & K. Shortridge, vol. 522 of ASP Conf. Ser., 545

Servillat, M., Aicardi, S., Cecconi, B., & Mancini, M. 2021, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici, vol. TBD of ASP Conf. Ser., 999 TBD

Servillat, M., Boisson, C., Lefaucheur, J., Kosack, K., Sanguillon, M., Louys, M., & Bonnarel, F. 2020a, in ADASS XXVII, edited by P. Ballester, J. Ibsen, M. Solar, & K. Shortridge, vol. 522 of ASP Conf. Ser., 199

Servillat, M., Riebe, K., Boisson, C., Bonnarel, F., Galkin, A., Louys, M., Nullmeier, M., Renault-Tinacci, N., Sanguillon, M., & Streicher, O. 2020b, IVOA Provenance Data Model Version 1.0, IVOA Recommendation 11 April 2020

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. 2016, Scientific Data, 3