International
Virtual
Observatory
Alliance

Last-step flat provenance metadata

Version 0.1

IVOA Note 2022-04-14

Working Group
DM
This version
https://www.ivoa.net/documents/LastStepProvenance/20220414
Latest version
https://www.ivoa.net/documents/LastStepProvenance
Previous versions
This is the first public release
Author(s)
Mathieu Servillat
Editor(s)
Mathieu Servillat

Abstract

We propose a simplified structure to describe the immediate provenance of an entity as a single record, based on the IVOA Provenance Data Model (Servillat and Riebe et al., 2020), in order to facilitate its use and adoption. We thus define a flat list of attributes to describe the last activity that led to the generation of an entity, in the particular case of digital dataset generation. Following the Provenance Data Model, this activity may be related to used entities and other entities generated by this activity. The last activity may be further characterized by configuration parameters, a software description and a context. The context may be the execution of a workflow (seen as a sequence of planned activities with a purpose), that may have used initially an instrument, e.g. to perform observations or the acquisition of raw data. Such a flat list of attributes could be stored in the header of a FITS file (keyword names proposed), as a file (YAML structure proposed), or in a database (possibly as a view on top of a provenance database following the IVOA Provenance Data Model).

Status of this document

This is an IVOA Note expressing suggestions from and opinions of the authors. It is intended to share best practices, possible approaches, or other perspectives on interoperability with the Virtual Observatory. It should not be referenced or otherwise interpreted as a standard specification.

A list of current IVOA Recommendations and other technical documents can be found at https://www.ivoa.net/documents/.

Contents

Acknowledgments

We acknowledge support from the ESCAPE project funded by the EU Horizon 2020 research and innovation program (Grant Agreement n.824064).

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).

Conformance-related definitions

The words "MUST", "SHALL", "SHOULD", "MAY", "RECOMMENDED", and "OPTIONAL" (in upper or lower case) used in this document are to be interpreted as described in IETF standard RFC2119 (Bradner, 1997).

The *Virtual Observatory (VO)* is a general term for a collection of federated resources that can be used to conduct astronomical research, education, and outreach. The International Virtual Observatory Alliance (IVOA) is a global collaboration of separately funded projects to develop standards and infrastructure that enable VO applications.

1 Introduction

Astronomical observatories and data providers are increasingly involved in the development of Open Science. The process of making data FAIR[1] (Findable, Accessible, Interoperable and Reusable) often has to be integrated early in the development of astronomical projects. Since more than 20 years, the IVOA[2] (International Virtual Observatory Alliance) provides various standards to foster interoperability and enable the production of FAIR data.

The Reusable principle is more subjective and requires rich metadata to demonstrate the quality, reliability and trustworthiness of the data. Detailed provenance is thus a key information to provide along with the astronomical data. The IVOA validated in April 2020 a Provenance Data Model (Servillat and Riebe et al., 2020) to structure this information. It is based on the W3C PROV concepts of Entity, Activity and Agent (Moreau and Missier et al., 2013) with a dedicated set of classes for activity description (e.g. method, algorithm, software) and activity configuration (e.g. parameters).

2 Diagram and main concepts

Full provenance graphs can become extremely complex. We propose to define a last-step provenance scheme as a single record, limited to the last activity (execution and software description) and the context (workflow, instrument), that may be embedded into an entity as a list of keywords.

The last-step provenance should include identifiers (of entities, activities and agents), in particular the identifiers of generated and used entities, so

[1] https://www.go-fair.org/fair-principles
[2] https://www.ivoa.net

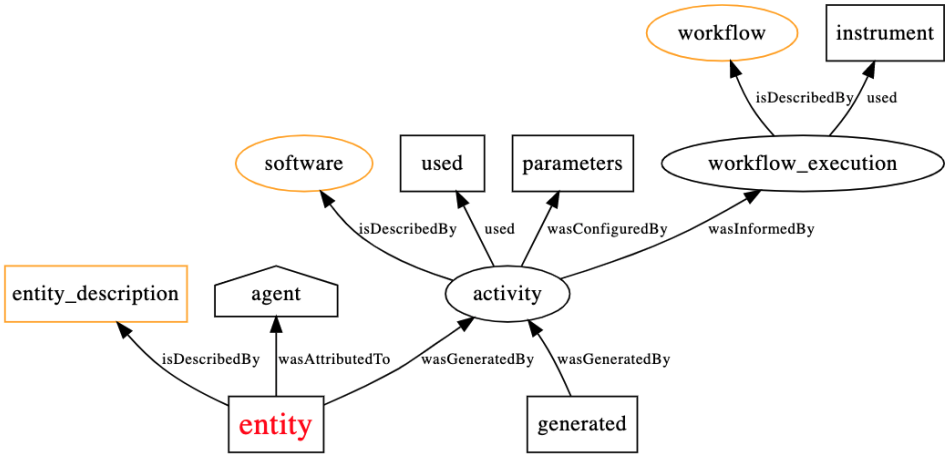

Figure 1: Last-step provenance diagram of an entity. Information related to entities are shown as rectangles, activities as ellipses, agents as house-shaped pentagons, and descriptions appear in orange. Relations are labelled as defined in the IVOA Provenance Data Model (Servillat and Riebe et al., 2020).

that a full provenance may be reconstructed from a sequence of last-step provenance records recursively.

The diagram of a last-step provenance record is shown in Figure 1 using classes defined in the IVOA Provenance Data Model, and the list of attributes is presented in Table 1 (where UTypes indicate the model components).

The following terms are used in the diagram:

- **entity**: the main entity generated, which is a digital dataset that may contain the last-step provenance record.
- **activity**: the last activity , i.e. the activity that generated the main entity.
- **used**, **generated**: entities related to the activity indicated by their identifiers.
- **entity_description**: description of the main entity
- **agent**: main contact responsible for the entity
- **software**: description of the activity, i.e. information on the software executed to generate the digital dataset.
- **parameters**: list of key-value pairs that configured the activity.

- **workflow**: sequence of activities planned to perform a process with a specific purpose. In the Provenance Data Model, this is seen as an Activity Description class.
- **workflow_execution**: super-activity corresponding to the execution of the workflow.
- **instrument**: the instrument that may have acquired the initial raw data that led to the generation of the main entity.

Table 1: List of attributes of a last-step provenance record

keyword	**UType**	**Description**
entity_id	Entity.id	
entity_location	Entity.location	
entity_generatedAtTime	Entity.generatedAtTime	
entity_comment	Entity.comment	
entity_name	EntityDescription.name	
entity_description	EntityDescription.description	
entity_type	EntityDescription.type	
entity_content_type	EntityDescription.content_type	
entity_docurl	EntityDescription.docurl	
agent_id	Agent.id	
agent_name	Agent.name	
agent_type	Agent.type	
agent_email	Agent.email	
activity_id	Activity.id	
activity_name	Activity.name	
activity_startTime	Activity.startTime	
activity_endTime	Activity.endTime	
activity_comment	Activity.comment	
activity_parameters	List of Parameter.name and Parameter.value	
used_ids	List of Entity.id	
generated_ids	List of Entity.id	
software_name	ActivityDescription.name	
software_version	ActivityDescription.version	
software_description	ActivityDescription.description	
software_type	ActivityDescription.type	
software_docurl	ActivityDescription.docurl	
workflow_id	Activity.id	
workflow_comment	Activity.comment	
workflow_name	ActivityDescription.name	
workflow_version	ActivityDescription.version	
workflow_description	ActivityDescription.description	
workflow_type	ActivityDescription.type	
workflow_docurl	ActivityDescription.docurl	
instrument_id	Entity.id	
instrument_location	Entity.location	
instrument_comment	Entity.comment	
instrument_name	EntityDescription.name	

keyword	UType	Description
instrument_description	EntityDescription.description	
instrument_type	EntityDescription.type	
instrument_docurl	EntityDescription.docurl	

Table 2: List of attributes with associated FITS keywords

keyword	FITS keyword	Alternative
entity_id	ENT_ID	
entity_location	ENT_LOC	
entity_generatedAtTime	ENT_GTIM	
entity_comment	ENT_COMM	
entity_name	ENT_NAME	
entity_description	ENT_DESC	
entity_type	ENT_TYPE	
entity_content_type	ENT_CTYP	
entity_docurl	ENT_DURL	
agent_id	AGT_ID	
agent_name	AGT_NAME	
agent_type	AGT_TYPE	
agent_email	AGT_MAIL	
activity_id	ACT_ID	
activity_name	ACT_NAME	
activity_startTime	ACT_STIM	
activity_endTime	ACT_ETIM	
activity_comment	ACT_COMM	
activity_parameters	PARN_001 to PARN_999	
	PARV_001 to PARV_999	
used_ids	USD_001 to USD_999	
generated_ids	GEN_001 to GEN_999	
software_name	SFW_NAME	
software_version	SFW_VERS	
software_description	SFW_DESC	
software_type	SFW_TYPE	
software_docurl	SFW_DURL	
workflow_id	WKF_ID	
workflow_comment	WKF_COMM	
workflow_name	WKF_NAME	
workflow_version	WKF_VERS	
workflow_description	WKF_DESC	
workflow_type	WKF_TYPE	
workflow_docurl	WKF_DURL	
instrument_id	INS_ID	
instrument_location	INS_LOC	
instrument_name	INS_NAME	
instrument_description	INS_DESC	
instrument_type	INS_TYPE	
instrument_docurl	INS_DURL	
instrument_comment	INS_COMM	

3 YAML Serialization

In addition to the concept of last-step provenance, we propose a structured serialization, easier to read by persons compared with W3C PROV formats.

The serialization of a last-step provenance record is presented in a YAML format, where each attribute is written <attribute> in the following example:

```yaml
agents:
   <agent_id>:
      name: <agent_name>
      type: <agent_type>
      email: <agent_email>
entities:
   <entity_id> :
      location: <entity_location>
      generatedAtTime: <entity_generatedAtTime>
      name: <entity_name>
      comment: <entity_comment>
      entity_description: <entity_name>
      attributed:
         - agent_id: <agent_id>
           role: Contact
   <instrument_id> :
      location: <instrument_location>
      name: <instrument_name>
      comment: <instrument_comment>
      entity_description: <instrument_name>
activities:
   <workflow_id>:
      comment: <workflow_comment>
      activity_description: <workflow_name>
   <activity_id>:
      name: <activity_name>
      startTime: <activity_startTime>
      endTime: <activity_endTime>
      activity_description: <activity_name>
      parameters:
         <name>: <value>  # from <activity_parameters>
         ...
      used:
         - entity_id: <used_id>  # from <used_ids>
         - entity_id: ...
         ...
```

```
      generated:
         - entity_id: <generated_id>  # from <generated_ids>
         - entity_id: ...
         ...
      informed:
         - activity_id: <workflow_id>
entity_descriptions:
   <entity_name>:
      description: <entity_description>
      type: <entity_type>
      content_type: <entity_content_type>
      docurl: <entity_docurl>
   <instrument_name>:
      description: <instrument_description>
      type: <instrument_type>
      docurl: <instrument_docurl>
activity_descriptions:
   <software_name>:
      version: <software_version>
      description: <software_description>
      type: <software_type>
      docurl: <software_docurl>
   <workflow_name>:
      version: <workflow_version>
      description: <workflow_description>
      type: <workflow_type>
      docurl: <workflow_docurl>
```

A Changes from Previous Versions

No previous versions yet.

References

Bradner, S. (1997), 'Key words for use in RFCs to indicate requirement levels', RFC 2119.
http://www.ietf.org/rfc/rfc2119.txt

Moreau, L., Missier, P., 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. and Tilmes, C. (2013), 'PROV-DM: The

prov data model', W3C Recommendation.
http://www.w3.org/TR/prov-dm

Servillat, M., Riebe, K., Boisson, C., Bonnarel, F., Galkin, A., Louys, M., Nullmeier, M., Renault-Tinacci, N., Sanguillon, M. and Streicher, O. (2020), 'IVOA Provenance Data Model Version 1.0', IVOA Recommendation 11 April 2020.
https://ui.adsabs.harvard.edu/abs/2020ivoa.spec.0411S