One-Step Provenance

IVOA Bologna 2023

Mathieu Servillat (LUTH - Observatoire de Paris / CNRS)
Catherine Boisson, François Bonnarel, Mireille Louys
+ ESCAPE participants
+ CTA members

IVOA Provenance

- **IVOA Provenance data model**
  - Follows W3C PROV
- **ProvSAP**: simple access protocole
  - Request provenance graph for an identifier (entity/activity)
  - Takes advantage of the W3C serializations (JSON, XML, SVG, PNG…)
  - `voprov` Python package
  - Implementations: Pollux, OPUS (CTA, MASER, CompOSE), ...
- **ProvTAP**: table access protocole
  - TAP Schema based on ProvenanceDM
  - Adding a few simplified views
  - Implementations : HiPS tiles provenance, …


International Virtual Observatory Alliance

IVOA Documents

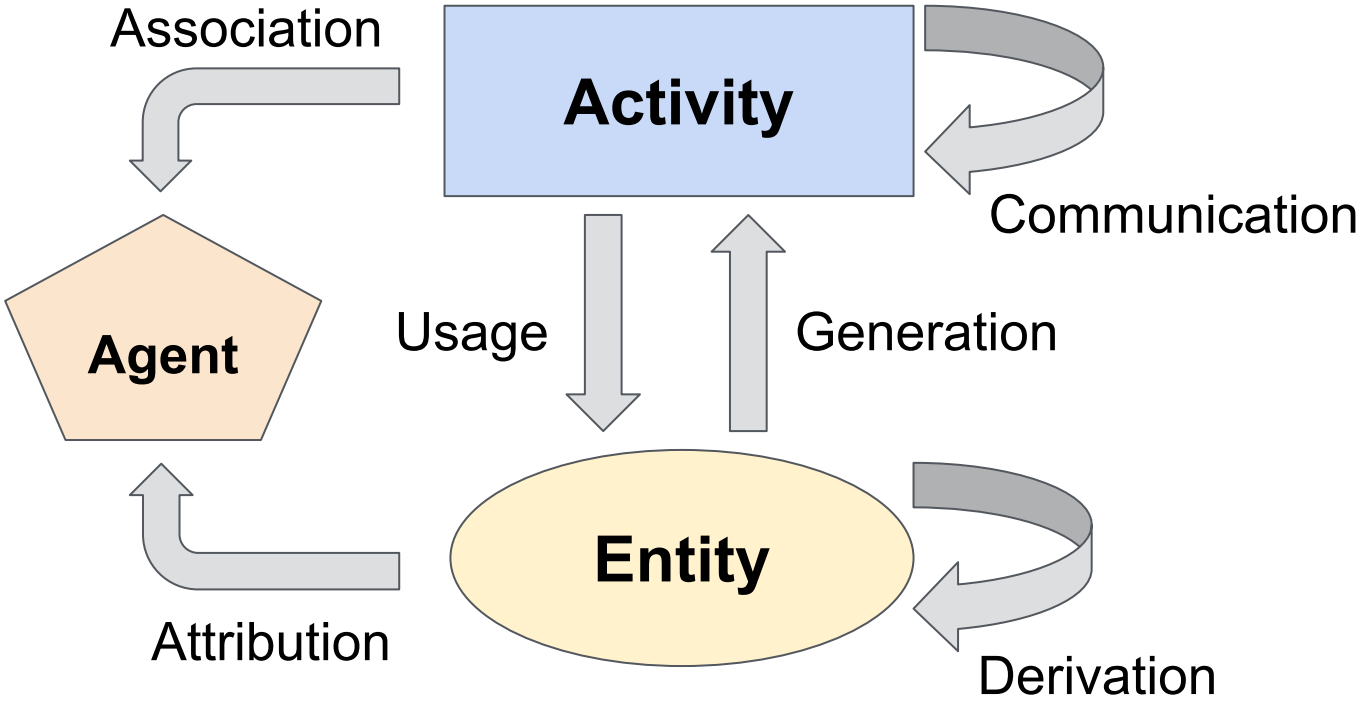
http://www.ivoa.net/documents/ProvenanceDM/

IVOA Provenance Data Model
Version 1.0

IVOA Recommendation 11 April 2020

Interest/Working Group:
http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel
Author(s):
Mathieu Servillat, Kristin Riebe, Catherine Boisson, François Bonnarel, Anastasia Galkin, Mireille Louys, Markus Nullmeier, Nicolas Renault-Tinacci, Michèle Sanguillon, Ole Streicher
Editor(s):
Mathieu Servillat

Provenance glossary

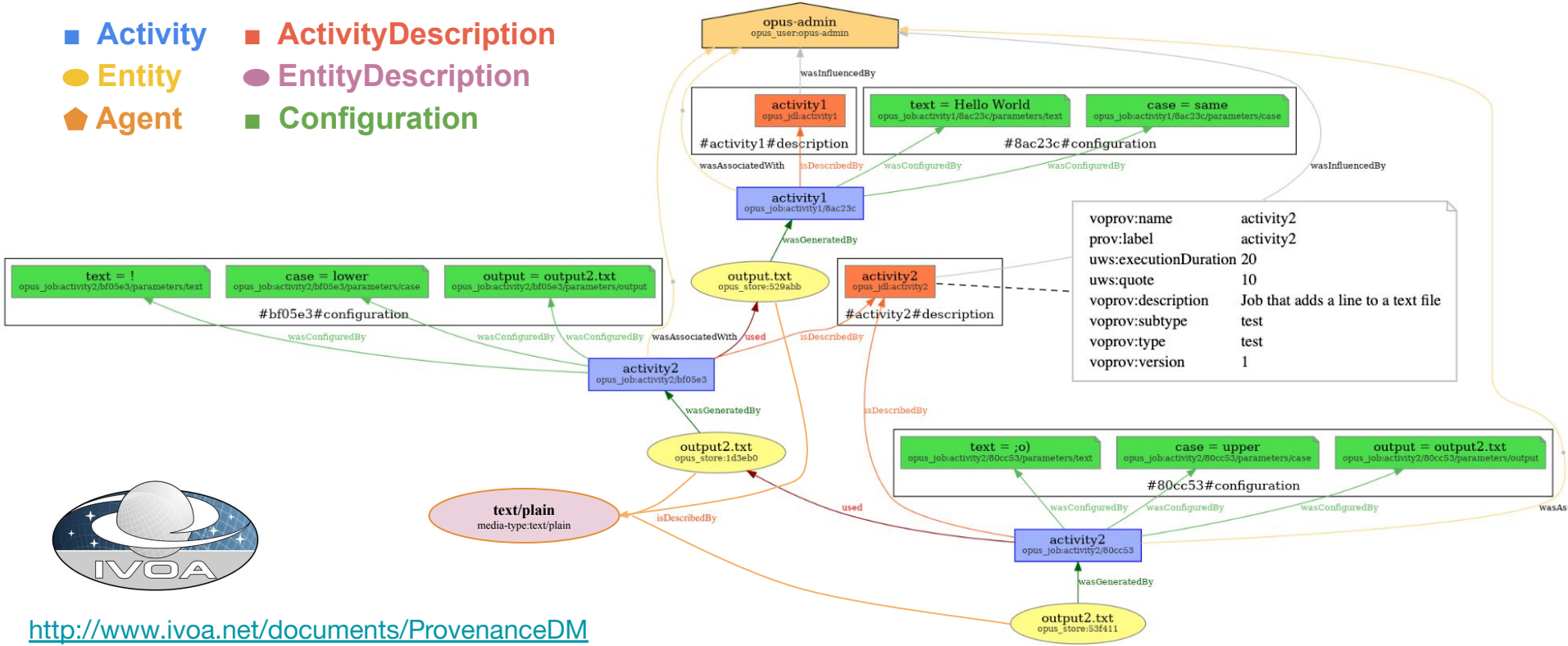


Word Wide Web Consortium

http://www.w3.org/TR/prov-overview

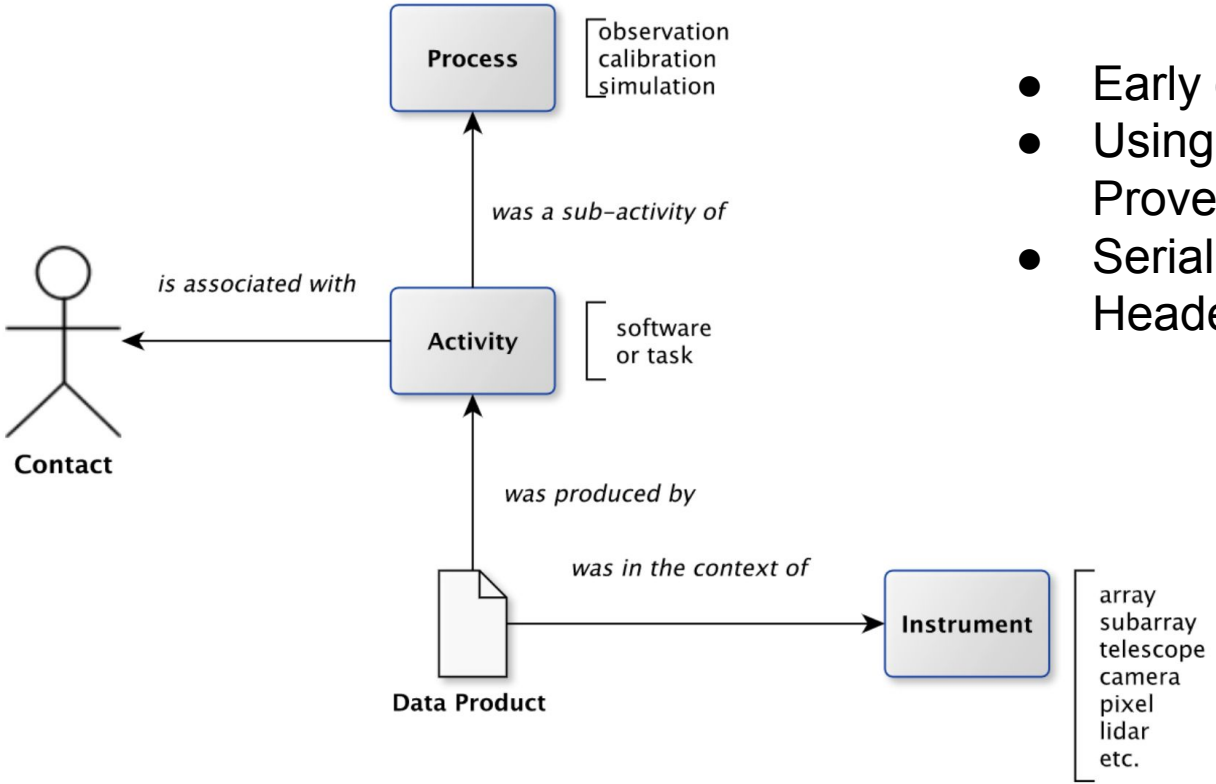
Full IVOA Provenance graph

- Activity
- ActivityDescription
- Entity
- EntityDescription
- Agent
- Configuration



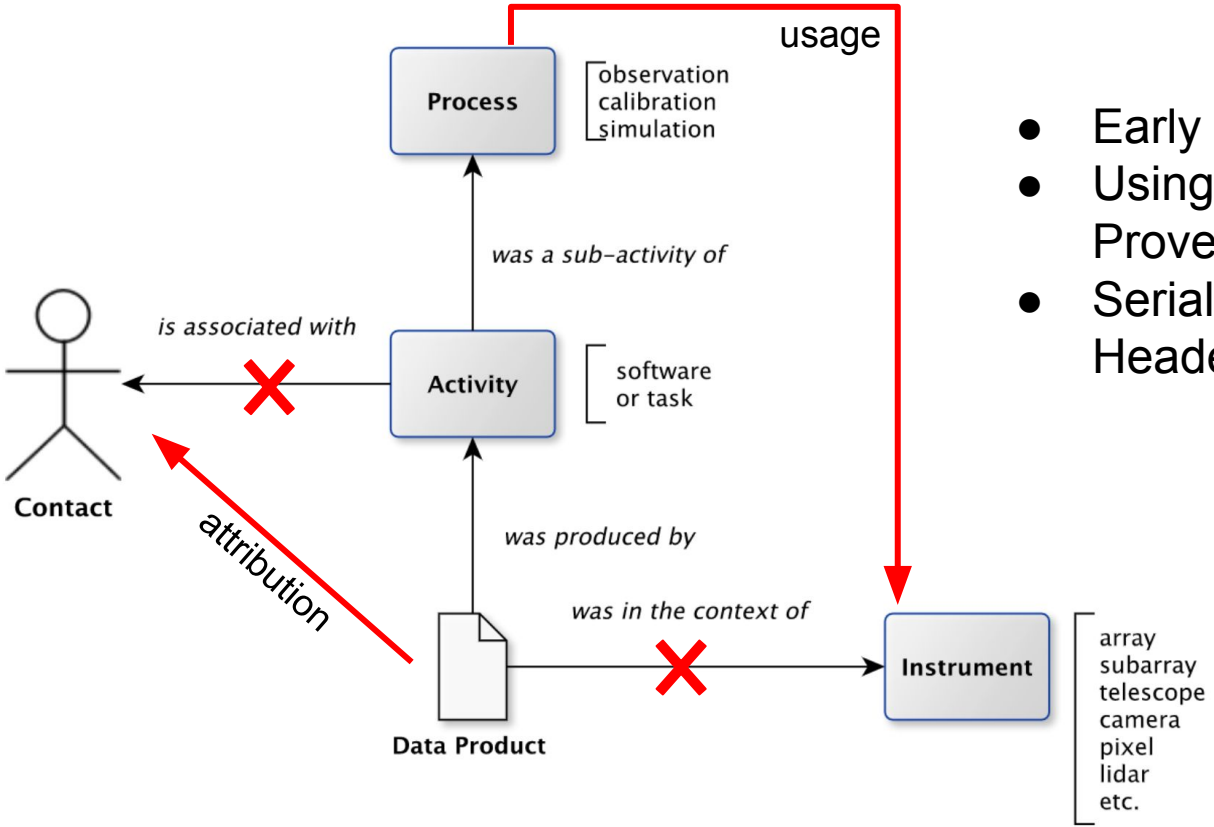
http://www.ivoa.net/documents/ProvenanceDM

CTAO Reference Metadata (draft)



- Early discussions in CTA
- Using some IVOA Provenance concepts
- Serialization in FITS Header

CTAO Reference Metadata



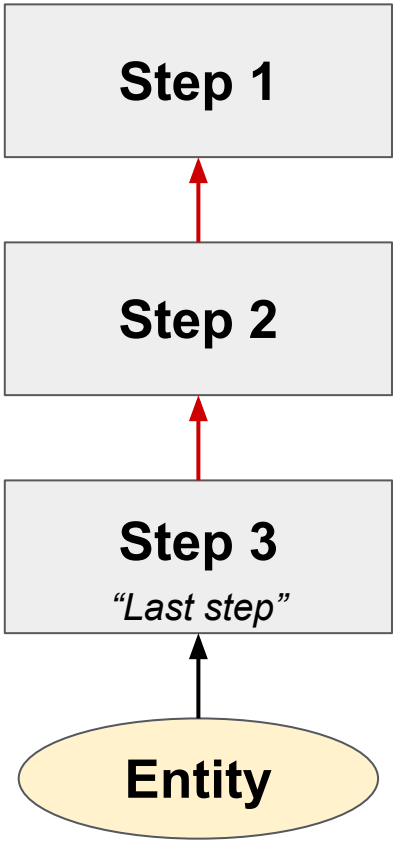
- Early discussions in CTA
- Using some IVOA Provenance concepts
- Serialization in FITS Header

Table 2.1 – Reference Metadata for CTA Observatory Data Products.

hierarch_0	hierarch_1	hierarch_n	example_1	example_2	example_3	example_4	description	required	type	enum	validation
CTA	REFERENCE	VERSION	1.0.0	1.0.0	1.0.0	1.0.0	version of the Reference metadata schema used in the data product	☑	string		="1.0.0"
CTA	CONTACT	ORGANIZATION	CTAO	LST Consortium	NectarCam Consortium	ASWG	Organization to which this data product is associated	☑	string		
CTA	CONTACT	NAME	CTAO Support	A. Scientist	A. Postdoc	D. Manager	Name of contact within organization	☑	string		
CTA	CONTACT	EMAIL	support@cta-observatory.org	a@b.com	a@b.com	a@b.com	Contact email address	☑	string		format: email
CTA	PRODUCT	DESCRIPTION	Cat C DL3 event list	Flatfielding Coefficients	Quantum Efficiency Curve	Proton Simulation	Human-readable description of data product	☑	string		
CTA	PRODUCT	CREATION_TIME	2018-11-10 15:30	2019-10-10 07:21	2018-11-14 18:03:21	2018-11-14 00:00	Human-readable date and time of file creation, in ISO format YYYY-MM-DD HH:MM:SS.ss, UTC	☐	string		format: datetime
CTA	PRODUCT	ID	*TBD*	*TBD*	*TBD*	*TBD*	a fixed-id to identify this product. E.g. a UUID or VFN.	☐	string		
CTA	PRODUCT	DATA CATEGORY	C	A	other	other	see CTA data identifier document	☑	string	A,B,C,S	
CTA	PRODUCT	DATA LEVEL	DL3	DL0	R0	R0	see CTA data identifier document	☑	string	R0, R1, DL0, DL1, DL2, DL3, DL4, DL5, DL6	
CTA	PRODUCT	DATA ASSOCIATION	Subarray	Telescope	Telescope	Site	see CTA data identifier document	☑	string	CTA, Site, Subarray, Telescope, Target	
CTA	PRODUCT	DATA TYPE	Event	Service	Service	Event	see CTA data identifier document	☐	string	Event, Monitoring, Service, DataCube, Catalog	

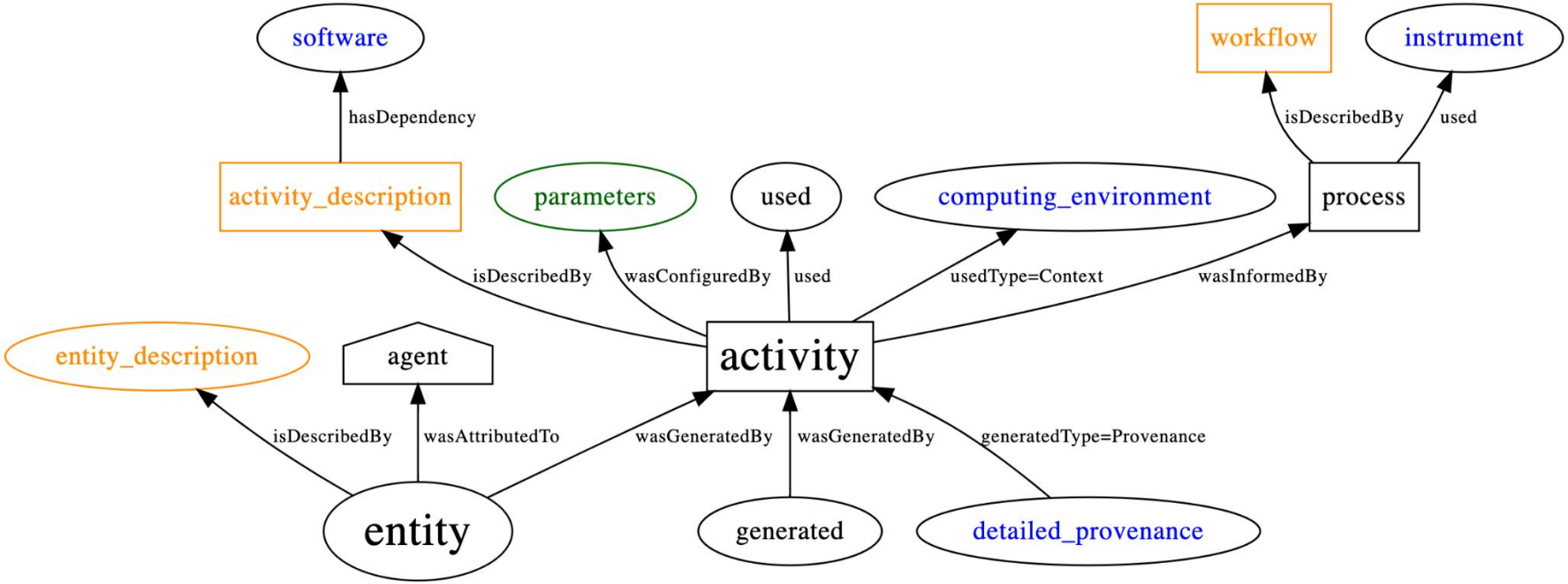
Back to IVOA Provenance

- Use cases
  - CTA data products header
  - Workshops with ESCAPE partners, within ASOV
  - ADASS BoF session
- Requirements
  - Cite **software**
  - Record the **context** of execution
  - Include provenance attributes **inside** an entity
  - Handle different levels of **details**
- Definition of a “One-Step Provenance”
  - **List of attributes** to describe one step of data generation
  - Links between steps using **identifiers**
  - The “Last step” may be embedded in the entity



One-Step Provenance Data Model

- **Subgraph** designed with IVOA Provenance concepts
- Applied to **digital object generation**
- Introduces **specialized entities** (software, computing environment, instrument)



Serialisation of a One-Step Provenance record

- Flat list of keywords
  - A few mandatory keywords (id)
  - other optional information to provide “detailed” provenance (see FAI**R** principles)
- Proposed FITS keywords
  - “Last step” provenance
  - embedded in the generated entity
- Identifiers are key
  - Resolved via a ProvSAP service

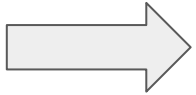
keyword	UType	FITS keyword
entity_id	Entity.id	ENT_ID
entity_location	Entity.location	ENT_LOC
entity_generatedAtTime	Entity.generatedAtTime	ENT_GTIM
entity_name	EntityDescription.name	ENT_NAME
entity_type	EntityDescription.type	ENT_TYPE
entity_content_type	EntityDescription.content_type	ENT_CTYP
entity_docurl	EntityDescription.docurl	ENT_DURL
entity_comment	Entity.comment	ENT_COMM
agent_id	Agent.id	AGT_ID
agent_name	Agent.name	AGT_NAME
agent_type	Agent.type	AGT_TYPE
agent_email	Agent.email	AGT_MAIL

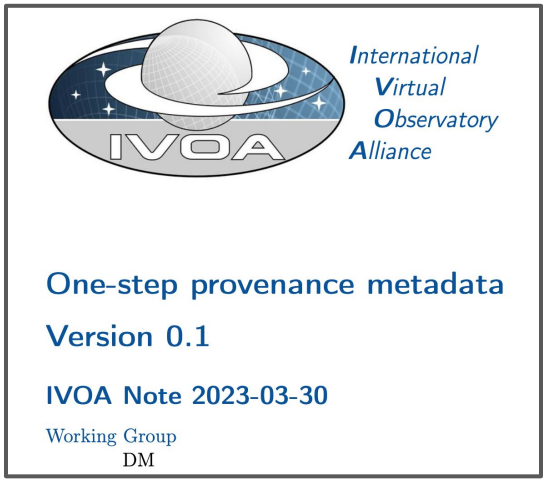
...

Specialized entities: context and links

- entityType = Instrument
  - name
  - identifier
  - but maybe more advanced model
- entityType = Software
  - name
  - version
  - identifier (e.g. doi)
- entityType = Computing Environment
  - Operating System
  - Nodes, memory…
  - Variables (path, python, …)

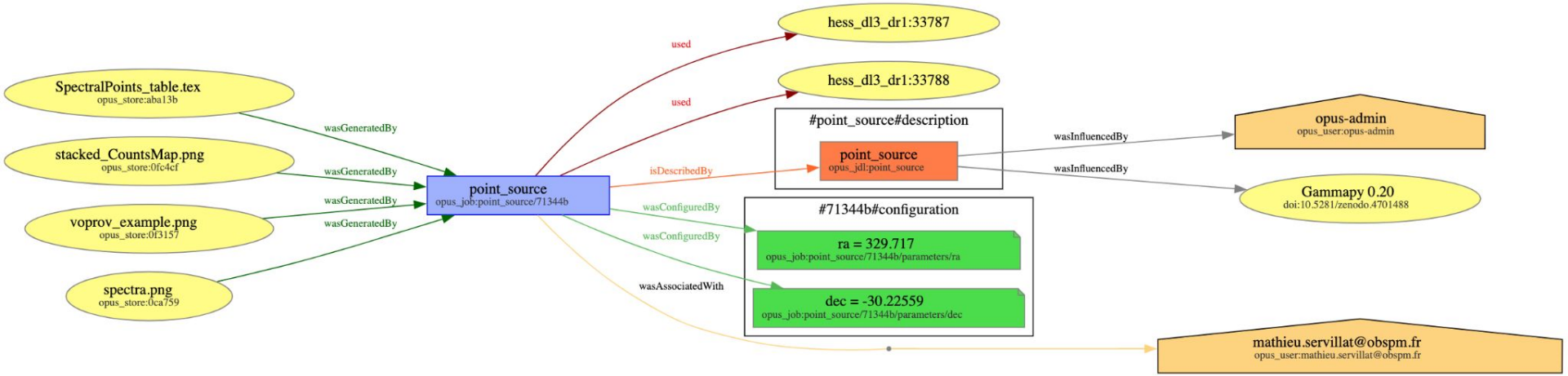
- Granularity of provenance
  - Link to more detailed provenance
- Link to Full Provenance
  - Identifiers should be resolved through a ProvSAP service





Example of implementation

- Analysis of Cherenkov data with gammapy using OPUS
  - OPUS: job manager based on UWS (https://opus-job-manager.readthedocs.io)
  - job definition = expected input/output + configuration parameters + description
- need to provide an **internal provenance** using the `voprov` Python package
  - datasets really used, software version really used…



Structured Serialisation of Provenance

- W3C formats
  - XML, JSON
  - structure by blocks
  - not easy to find an information
  - “storage only”
- Proposed YAML format
  - machine **and** human readable
  - include relations with entities/activities
  - Makes it easy to **manipulate** provenance

```
activity:
   <activity_id>:
      name: <activity_name>
      startTime: <activity_startTime>
      endTime: <activity_endTime>
      activity_description: <activity_name>
      parameters:
         <name>: <value>  # from <activity_parameters>
         ...
      used:
         - <used_id>  # from <used_ids>
         - ...
      generated:
         - <generated_id>  # from <generated_ids>
         - ...
      informed:
         - <workflow_id>
   <workflow_id>:
      comment: <workflow_comment>
      activity_description: <workflow_name>
...
```