Provenance for data providers

ESCAPE EDP Forum 2021-11

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

FAIR principles for data sharing

https://www.go-fair.org/fair-principles

FINDABLE

Unique identifiers and metadata are used to allow data to be located quickly and efficiently

ACCESSIBLE

Data is open, free and universally available for research discovery efforts

INTER-OPERABLE

A common programming language is used to allow use in a broad range of applications

REUSABLE

All data is clearly described and outlines associated data-use standards

From F-A-I to FAIR

→ **ADASS XXXI**
talk I4-001 - “FAIR standards for astronomical data” - S. O’Toole
talk O4-002 - “FAIR high level data for Cherenkov astronomy” - M. Servillat

Findable
Accessible
Interoperable
Reusable?

- **Findable-Accessible-Interoperable**
  - Use the **Virtual Observatory standards**, protocols and services
  - Define community **standards** where required
  - To be discussed early in projects, but **technical solutions exist**

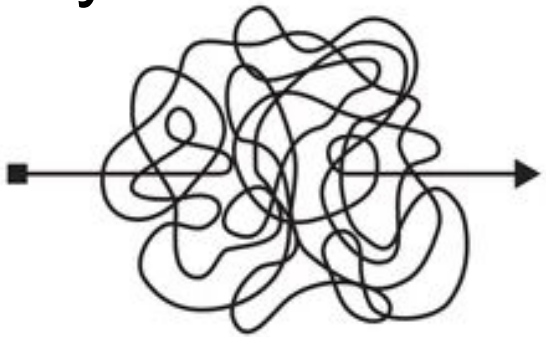
- **Reusability?**
  - Based on **trust,** need to prove the **quality** / **reliability** of the products

- **Reproducibility**
  - **A totally different goal**
  - Reproducible data may still be difficult to trust (if produced by a "black box")
  - Reusable data is not always automatically reproducible

- **What matters?**
  - **Tools** and **methods** used at each step of the process (e.g. software)
  - **How** it was executed (e.g. configuration parameters)
  - The **chain** of steps
  - **Sustainability**: with time, key information may disappear…

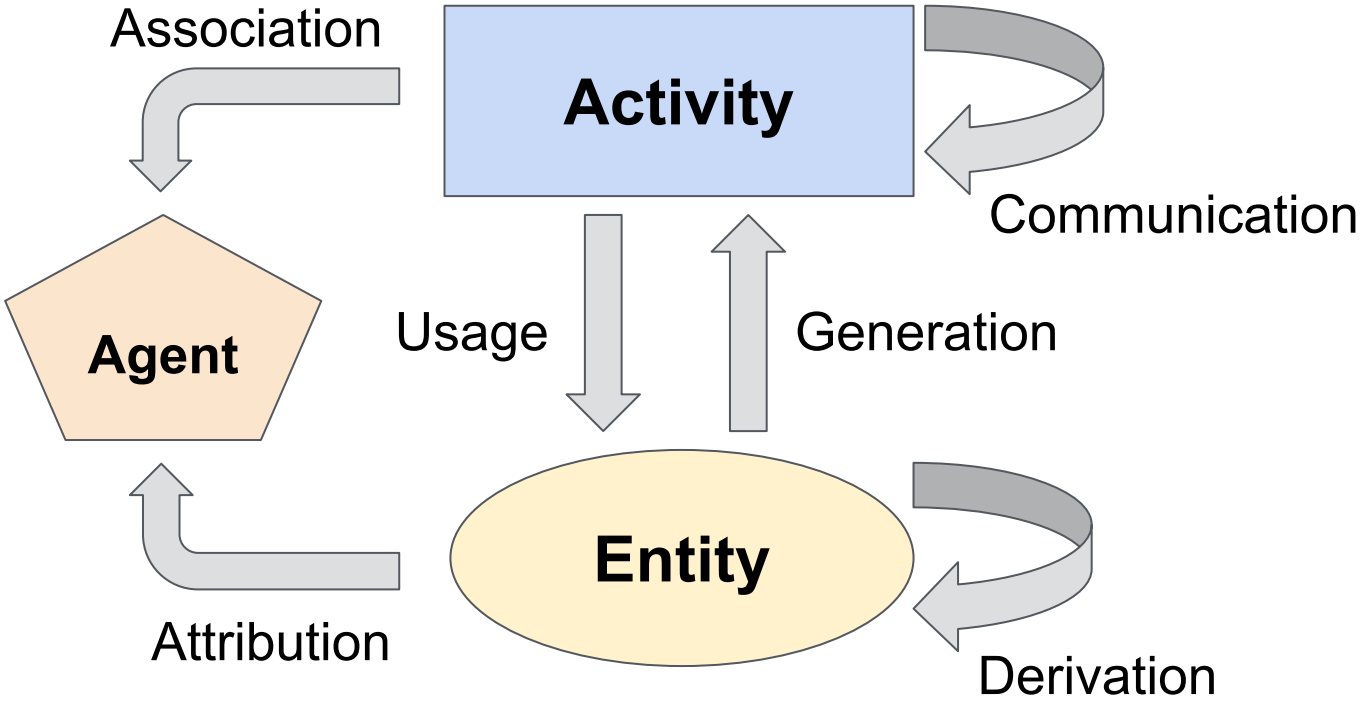
Provenance metadata

Provenance as an answer to reusability



- **Information to be recorded**
  - **origin** (generally not forgotten)
  - **+ path** (generally not *detailed* enough or *structured*)
- **How to record it?**
  - Keep the **trace** of what was used and generated at each step (easy)
  - **Identify** generated entities so that they can be *recognised* when used elsewhere (difficult!)
  - **Locate** and **describe** entities and activities
- **Store** provenance
  - in a central database following the data model
- **Access** to provenance
  - ProvSAP (Simple Access Protocol): extract a provenance graph
  - ProvTAP (Table Access Protocol): precise query on provenance metadata

Provenance glossary

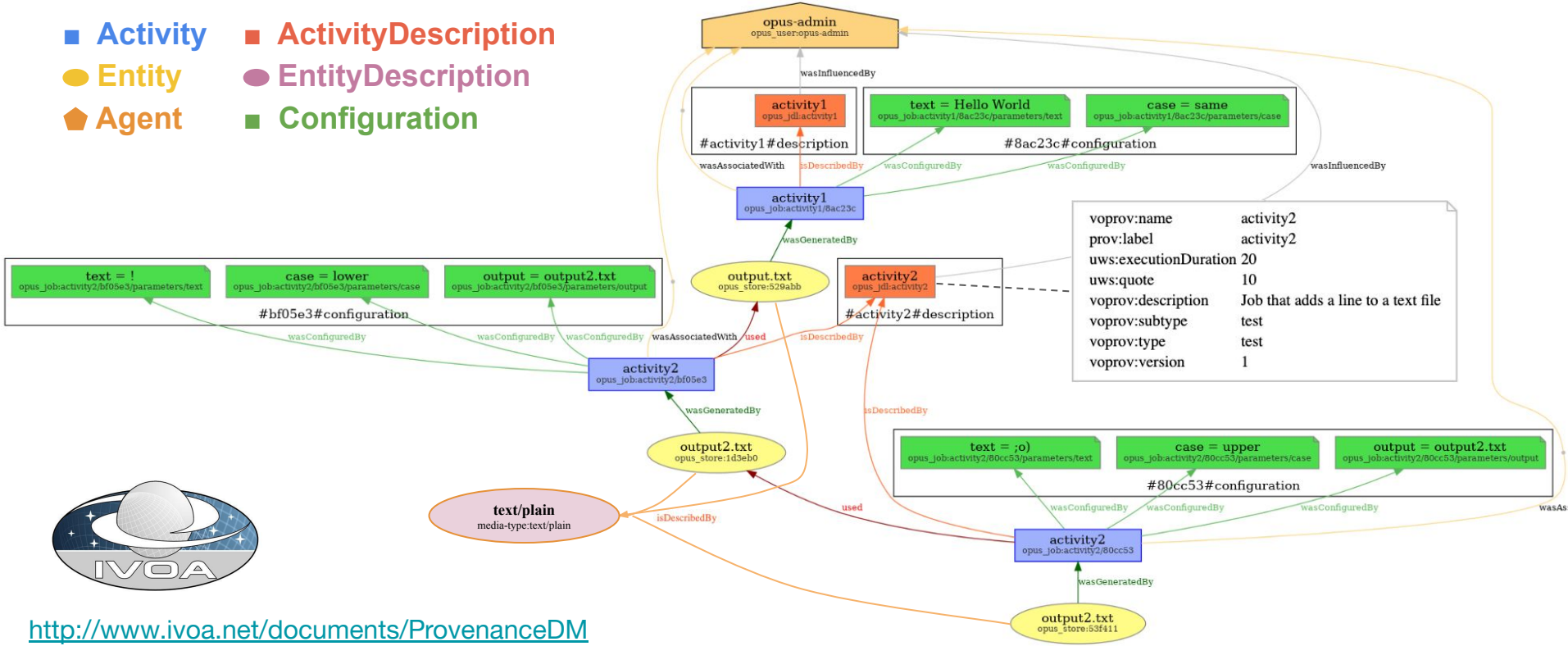


Word Wide Web Consortium

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

Full IVOA Provenance graph

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



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

A provenance management system

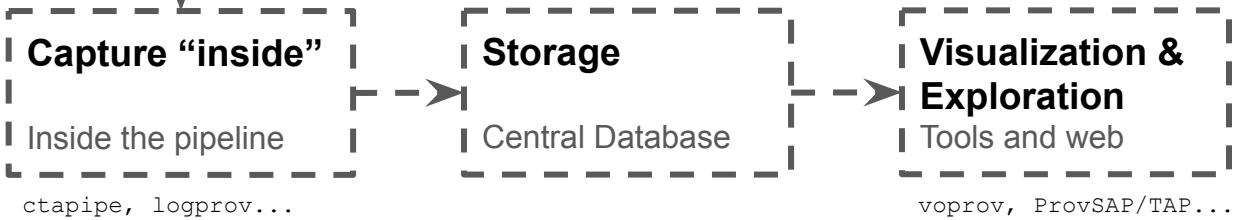
- What scientists generally have in mind:

Provenance Week 2021 proceedings : https://arxiv.org/abs/2109.07751

A provenance management system

- What scientists generally have in mind:

- But need for advanced provenance management:



Provenance Week 2021 proceedings : https://arxiv.org/abs/2109.07751

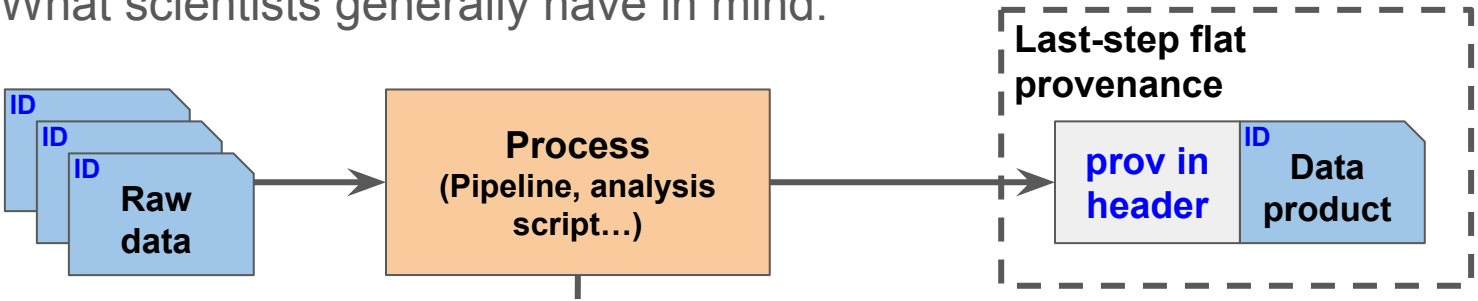
Some terminology

- **full provenance**: graph/tree/chain that **traces** activities and entities up to the raw data. This information is not hosted by the entities themselves, it should be stored in a central database, or as separate files.
- **end-user/specific "provenance"**: can be embedded into an entity, keywords or data that provides project specific **key information to use/analyse** the entity (e.g. for CTA: event class/type, telescope configuration, sky conditions, reco method...)
- **last-step provenance**: embedded into an entity as a list of keywords that gives some context and info on **last activity** (general workflow, software, versions, contact…), including the list of generated and used entity ids, so that a full provenance may be reconstructed from this minimum provenance.

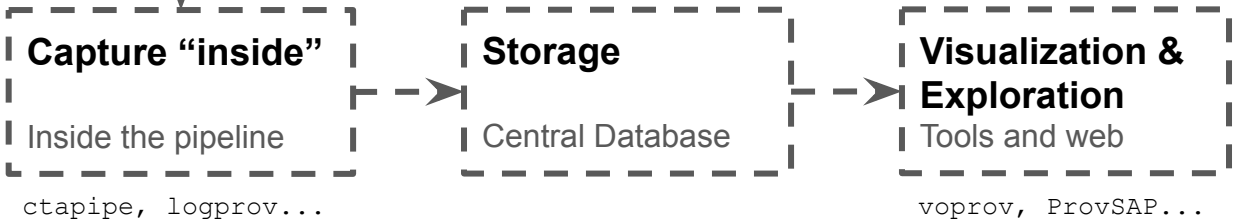
See ADASS XXX BoF proceedings : https://arxiv.org/abs/2101.08691
ESCAPE workshop on provenance : https://indico.in2p3.fr/event/21913/page/2641-summary

A provenance management system

- What scientists generally have in mind:



- But need for advanced provenance management:



Provenance Week 2021 proceedings : https://arxiv.org/abs/2109.07751

Last-step flat provenance

→ IVOA 2021-11 presentation
https://wiki.ivoa.net/internal/IVOA/InterOpNov2021DM/2021-11-04_Last-step_provenance_IVOA.pdf

- Problematic
  - Provenance graphs are complex, cannot be embedded in entities
  - Is there a **minimum** provenance?
  - Can provenance be expressed as a **flat** table?
  - Can provenance be **embedded**?
- Use cases
  - Workshop with ESCAPE partners
  - CTA data products header
- Content
  - 1 chain link
  - subgraph
  - keyword list
  - FITS keywords