Provenance activities in the European project ESCAPE

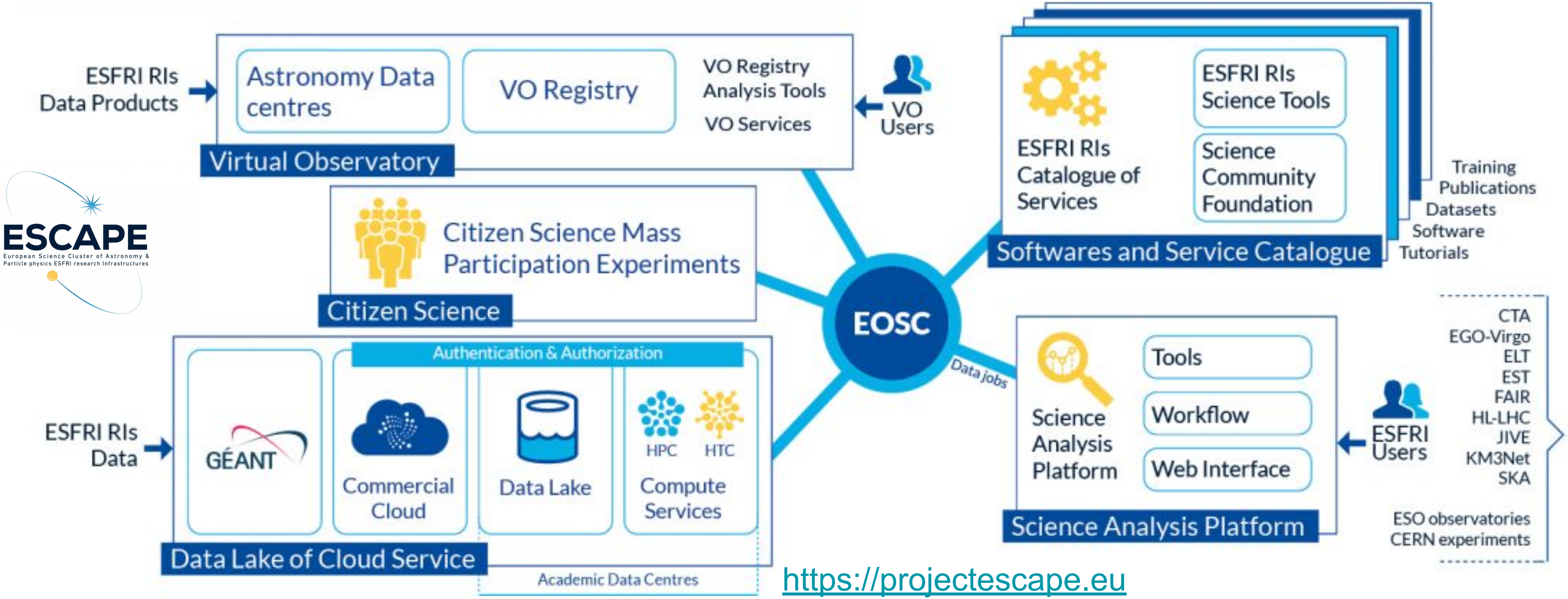
ESCAPE

European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

IVOA

EOSC – The **E**uropean **O**pen **S**cience **C**loud is a cloud for research data in Europe allowing for universal access to data;

ESCAPE – « The **E**uropean **S**cience **C**luster of **A**stronomy & **P**article Physics **E**SFRI Research Infrastructures » answers the EOSC ambition in bringing People, Data, Services, Training, Publications, Projects & Organisations, all together in an integrated and federated environment.



https://projectescape.eu

Provenance 2-days workshop on use cases

The objective was to collect the requirements of ESFRI projects in order to build the road map of future developments concerning provenance.

https://indico.in2p3.fr/event/21913/page/2641-summary

- Presentation of the **model**, and the associated implementations on provenance **capture**, **storage**, **access** and **visualization**
- Presentation of **each project use case**, following the pattern:
  - Project, context, data products
  - What is the relevant provenance information kept (or to be kept)?
  - How is it kept (or will be)?
  - What provenance will/should the end user see?
- General **discussion** on questions raised during the workshop

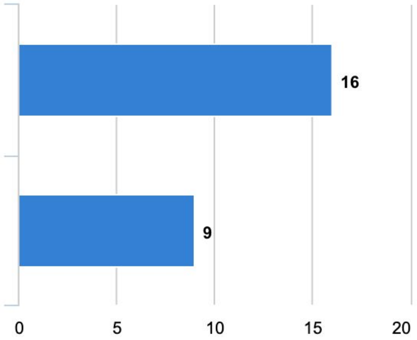
BoF at ADASS XXX on practical provenance

- **Introduction** to IVOA provenance
- **Questionnaire** posted and filled by attendants and others before
- **3 posters** : CTADIRAC (P9-250), VizieR (P9-216), OPUS (P9-89)
- **Discussion** (not really lively despite interest in the questionnaire...)
    - Minimum provenance: last step
    - Serialization / format: YAML, JSON are prefered ? VOTABLE (with model mapping)
    - Reproducibility can use provenance information:
        - Can we translate data flow language like CWL to IVOA provenance and reverse
        - Software parameters values may be stored with activities
        - Find out datasets produced with a given version of software
        - Proposal to use code ID to identify software in ActivityDescription
    - Visualisation using voprov library (see posters)
    - Access using VO protocols

Some answers to the questionnaire

1/ Are you a provider or a user of provenance information (or both)?

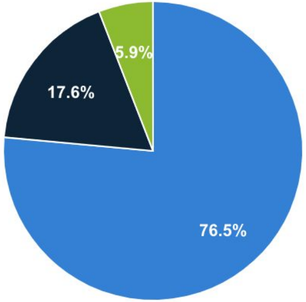
Chart options »



provider	16
user	9

2/ Do you already attach provenance information to the data you create or provide?

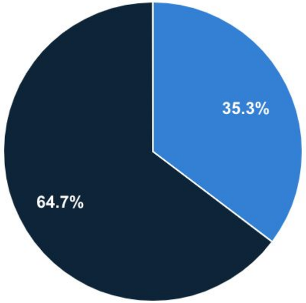
Chart options »



yes	13
no	3
not applicable	1

3/ Do you use the IVOA Provenance data model ?

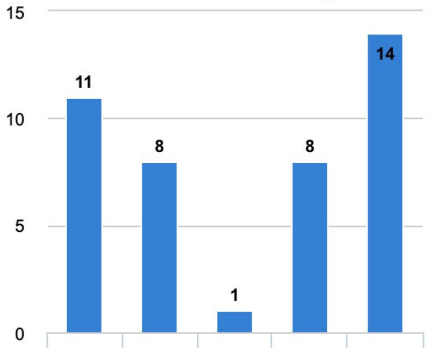
Chart options »



yes	6
no	11

5/ What is the most important goal in relation to provenance for you?

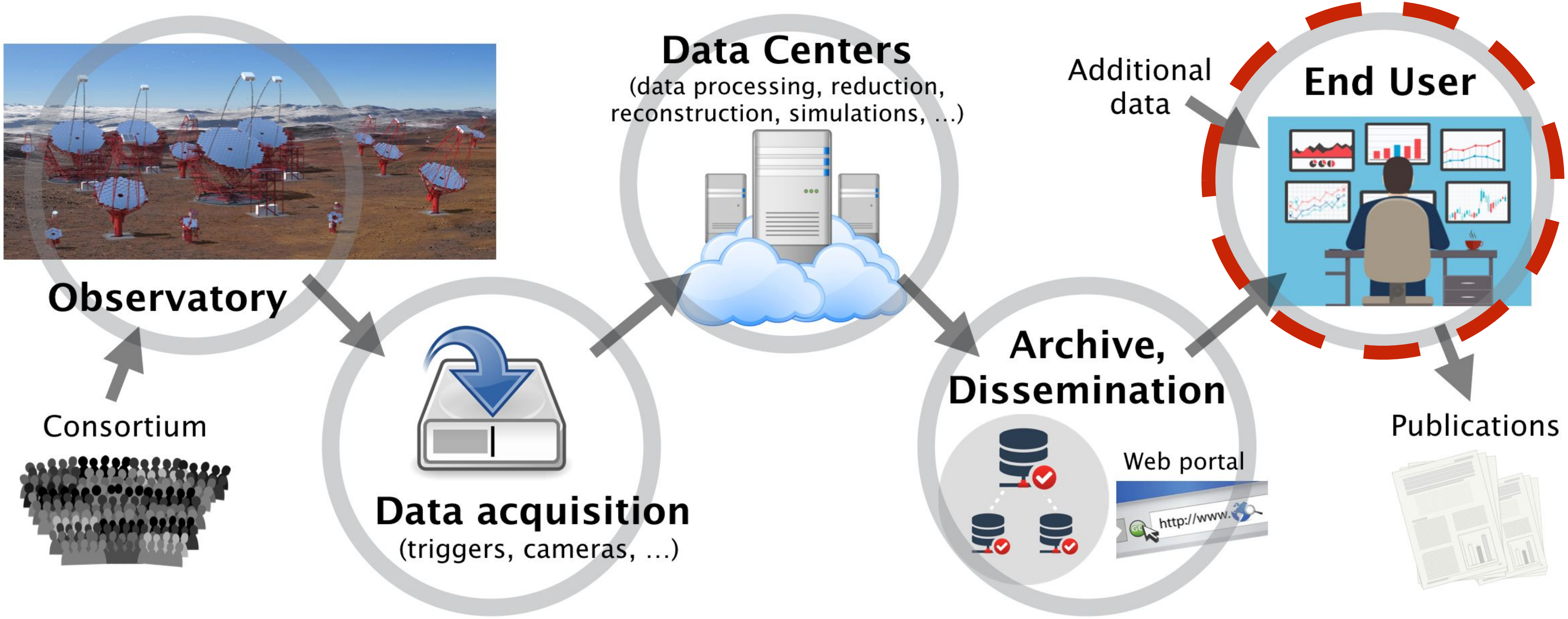
Chart options »



Traceability	11
Quality / Reliability	8
Acknowledgement	1
Debugging	8
Reproducibility	14

Astronomy context for data generation

Data product generation **obscure** to end user



Need **structured** and **detailed** provenance information!

FAIR principles for data sharing

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

Findable

F1. (Meta)data are assigned a globally unique and persistent **identifier**

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly include the **identifier** of the data they describe

F4. (Meta)data are **registered** or **indexed** in a searchable resource

Accessible

A1. (Meta)data are retrievable by their **identifier** using a **standardised** communications **protocol**

A1.1. The **protocol** is open, free, and universally implementable

A1.2. The **protocol** allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation.

I2. (Meta)data use **vocabularies** that follow FAIR principles

I3. (Meta)data include **qualified** references to other (meta)data

Reusable (+ Reproducible?)

R1. (Meta)data are richly described with a plurality of **accurate** and **relevant** attributes

R1.1. (Meta)data are released with a **clear** and accessible data usage **license**

R1.2. (Meta)data are associated with detailed **provenance**

R1.3. (Meta)data meet domain-relevant community **standards**

Why recording structured provenance?

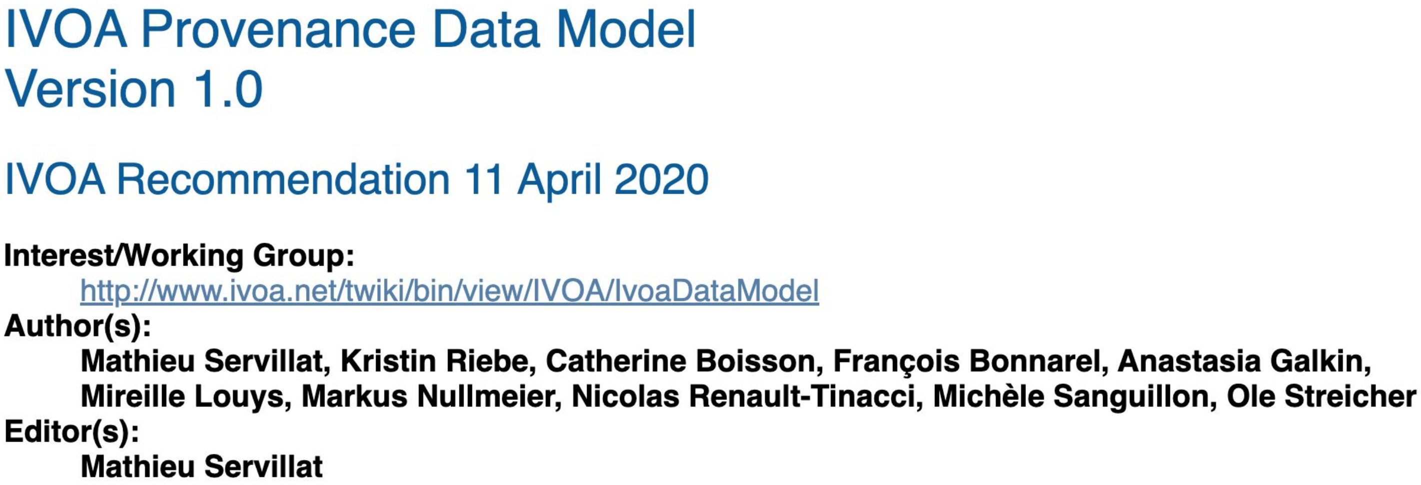
- **FAIR principles** (Findable, Accessible, Interoperable, Reusable)
  - https://www.go-fair.org/fair-principles/
  - "**rich**" metadata, following standard data model, protocols and formats
  - "**detailed provenance**"
- **Quality** / **Reliability** / **Trustworthiness** of the products
- **Reproducibility requirement** in projects
  - Be able to rerun each activity (maybe testing and improving each step)
  - Not necessary to keep every intermediate file that is easily reproducible (possible gain on disk space and costs)
- **Debugging**
  - Not necessary to restart from scratch: locate in the provenance tree the faulty parts or the products to be discarded

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

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

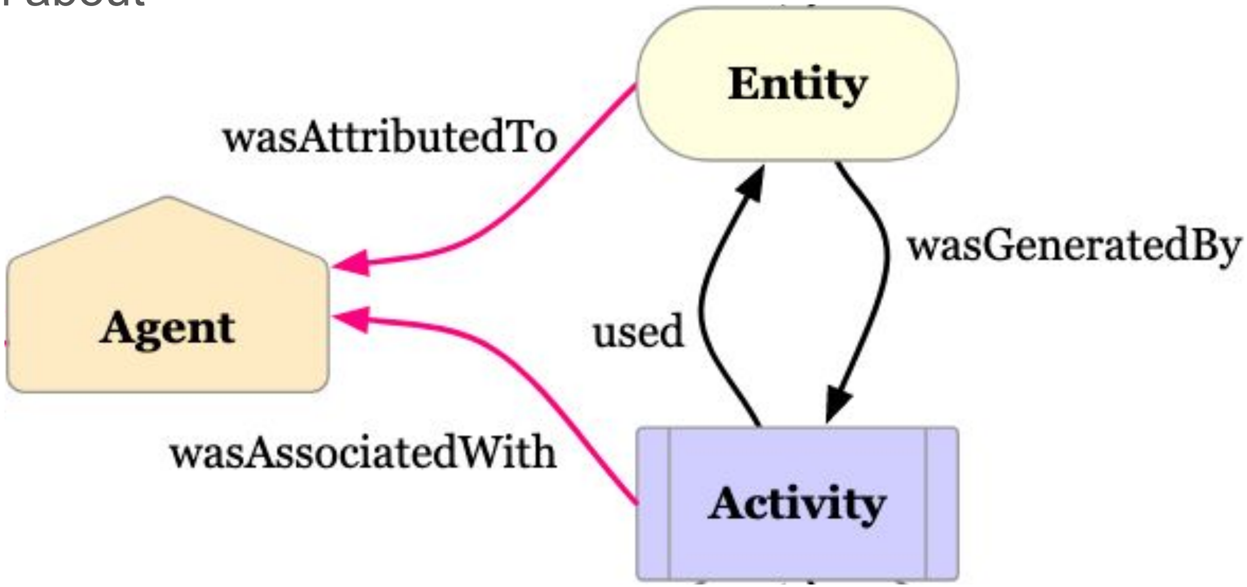
http://www.w3.org/TR/prov-overview/

W3C Provenance definition

Word Wide Web Consortium

W3C PROV (PROV-DM, 2013)

Provenance is 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**.

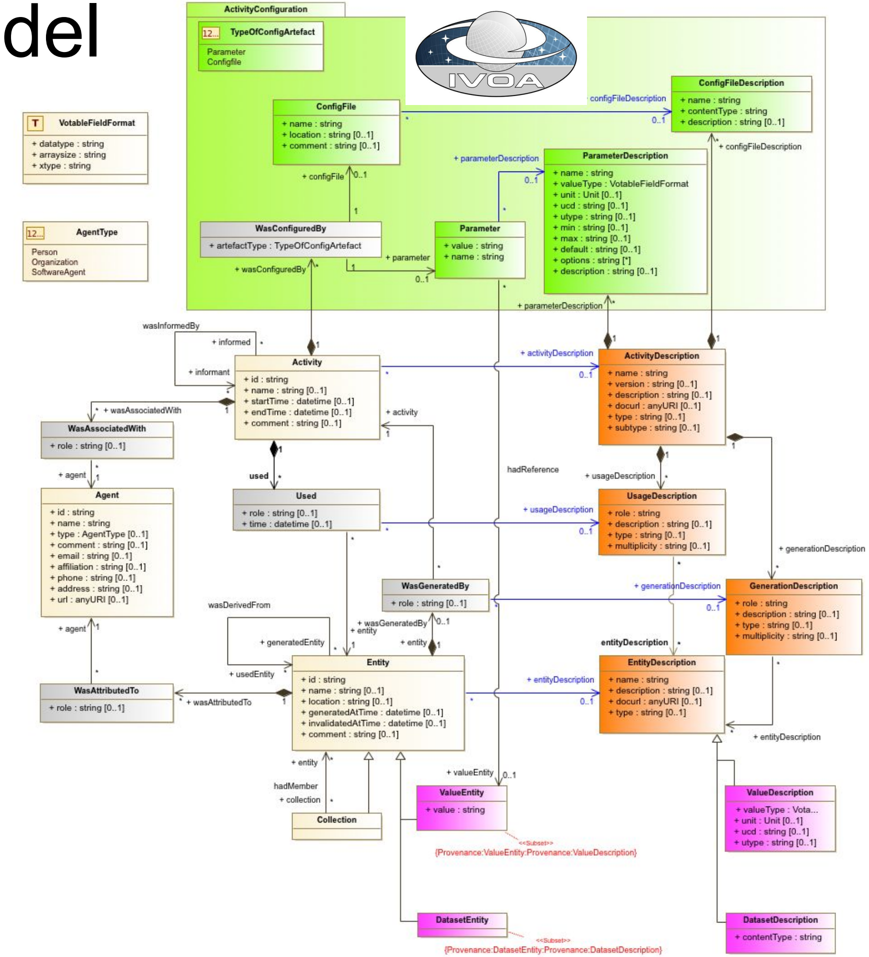


IVOA Provenance Data Model

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

Recommendation 1.0 in April 2020

- Adds “**Description**” classes
- Adds “**Configuration**” classes
- Plugged in with
  - **VO** data models and concepts (UCD, VOUnit, VOTable…)
  - **VO** access protocols (ProvTAP, ProvSAP)
- Serializations
  - W3C PROV (XML, JSON, SVG…)
  - **VO** specific (VOTable)



ActivityConfiguration

TypeOfConfigArtefact
Parameter
Configfile

ConfigFile
+ name : string
+ location : string [0..1]
+ comment : string [0..1]

ConfigFileDescription
+ name : string
+ contentType : string
+ description : string [0..1]

VotableFieldFormat
+ datatype : string
+ arraysize : string
+ xtype : string

AgentType
Person
Organization
SoftwareAgent

WasConfiguredBy
+ artefactType : TypeOfConfigArtefact

Parameter
+ value : string
+ name : string

ParameterDescription
+ name : string
+ valueType : VotableFieldFormat
+ unit : Unit [0..1]
+ ucd : string [0..1]
+ utype : string [0..1]
+ min : string [0..1]
+ max : string [0..1]
+ default : string [0..1]
+ options : string [*]
+ description : string [0..1]

Activity
+ id : string
+ name : string [0..1]
+ startTime : datetime [0..1]
+ endTime : datetime [0..1]
+ comment : string [0..1]

ActivityDescription
+ name : string
+ version : string [0..1]
+ description : string [0..1]
+ docurl : anyURI [0..1]
+ type : string [0..1]
+ subtype : string [0..1]

WasAssociatedWith
+ role : string [0..1]

Agent
+ id : string
+ name : string
+ type : AgentType [0..1]
+ comment : string [0..1]
+ email : string [0..1]
+ affiliation : string [0..1]
+ phone : string [0..1]
+ address : string [0..1]
+ url : anyURI [0..1]

Used
+ role : string [0..1]
+ time : datetime [0..1]

UsageDescription
+ role : string
+ description : string [0..1]
+ type : string [0..1]
+ multiplicity : string [0..1]

WasGeneratedBy
+ role : string [0..1]

GenerationDescription
+ role : string
+ description : string [0..1]
+ type : string [0..1]
+ multiplicity : string [0..1]

WasAttributedTo
+ role : string [0..1]

Entity
+ id : string
+ name : string [0..1]
+ location : string [0..1]
+ generatedAtTime : datetime [0..1]
+ invalidatedAtTime : datetime [0..1]
+ comment : string [0..1]

EntityDescription
+ name : string
+ description : string [0..1]
+ docurl : anyURI [0..1]
+ type : string [0..1]

Collection

ValueEntity
+ value : string

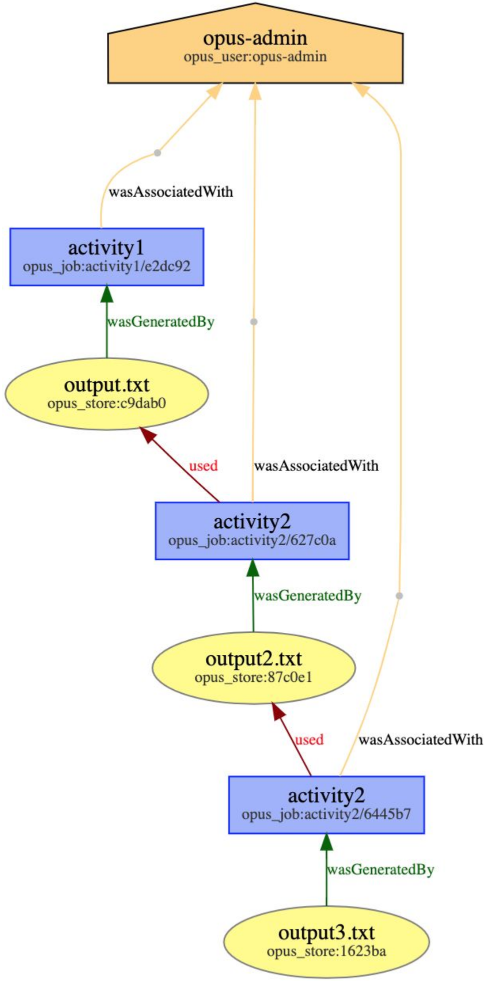
ValueDescription
+ valueType : Vota...
+ unit : Unit [0..1]
+ ucd : string [0..1]
+ utype : string [0..1]

DatasetEntity

DatasetDescription
+ contentType : string

<<Subset>>
{Provenance:ValueEntity:Provenance:ValueDescription}

<<Subset>>
{Provenance:DatasetEntity:Provenance:DatasetDescription}



Provenance graph

Provenance is :

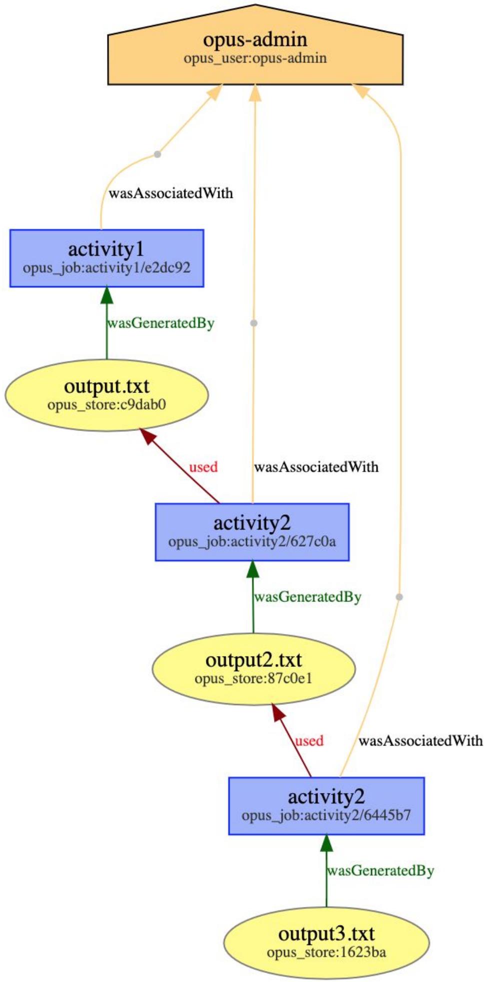
- a **chain** of activities and entities (used and generated)
- that occured in the **past**

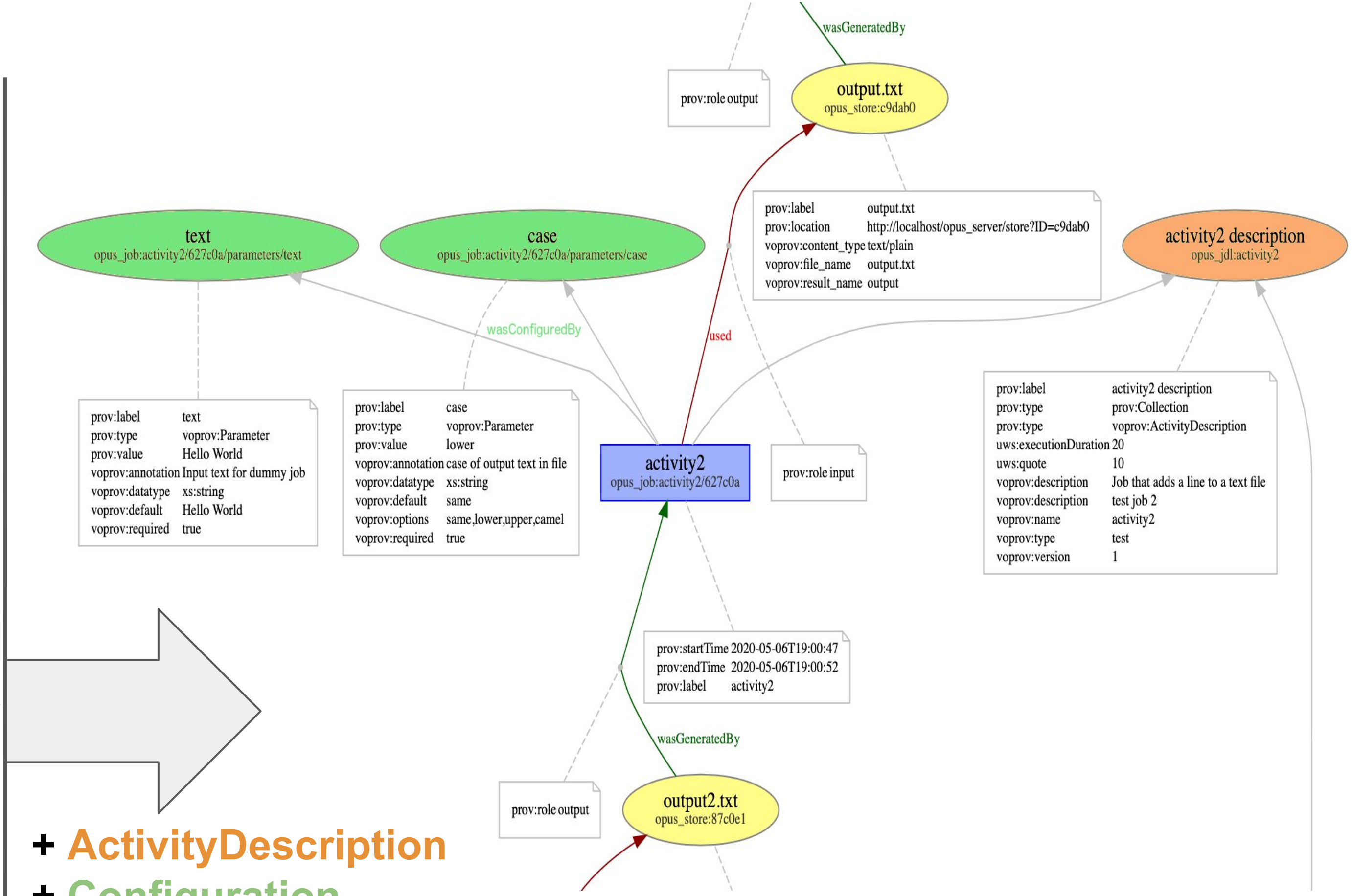
Using the **core data model**, some goals are achieved:

- Unique identifiers
- Traceability
- Contact and Acknowledgement

By using the **full IVOA data model**, more 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** ?





prov:label	output.txt
prov:location	http://localhost/opus_server/store?ID=c9dab0
voprov:content_type	text/plain
voprov:file_name	output.txt
voprov:result_name	output

prov:label	text
prov:type	voprov:Parameter
prov:value	Hello World
voprov:annotation	Input text for dummy job
voprov:datatype	xs:string
voprov:default	Hello World
voprov:required	true

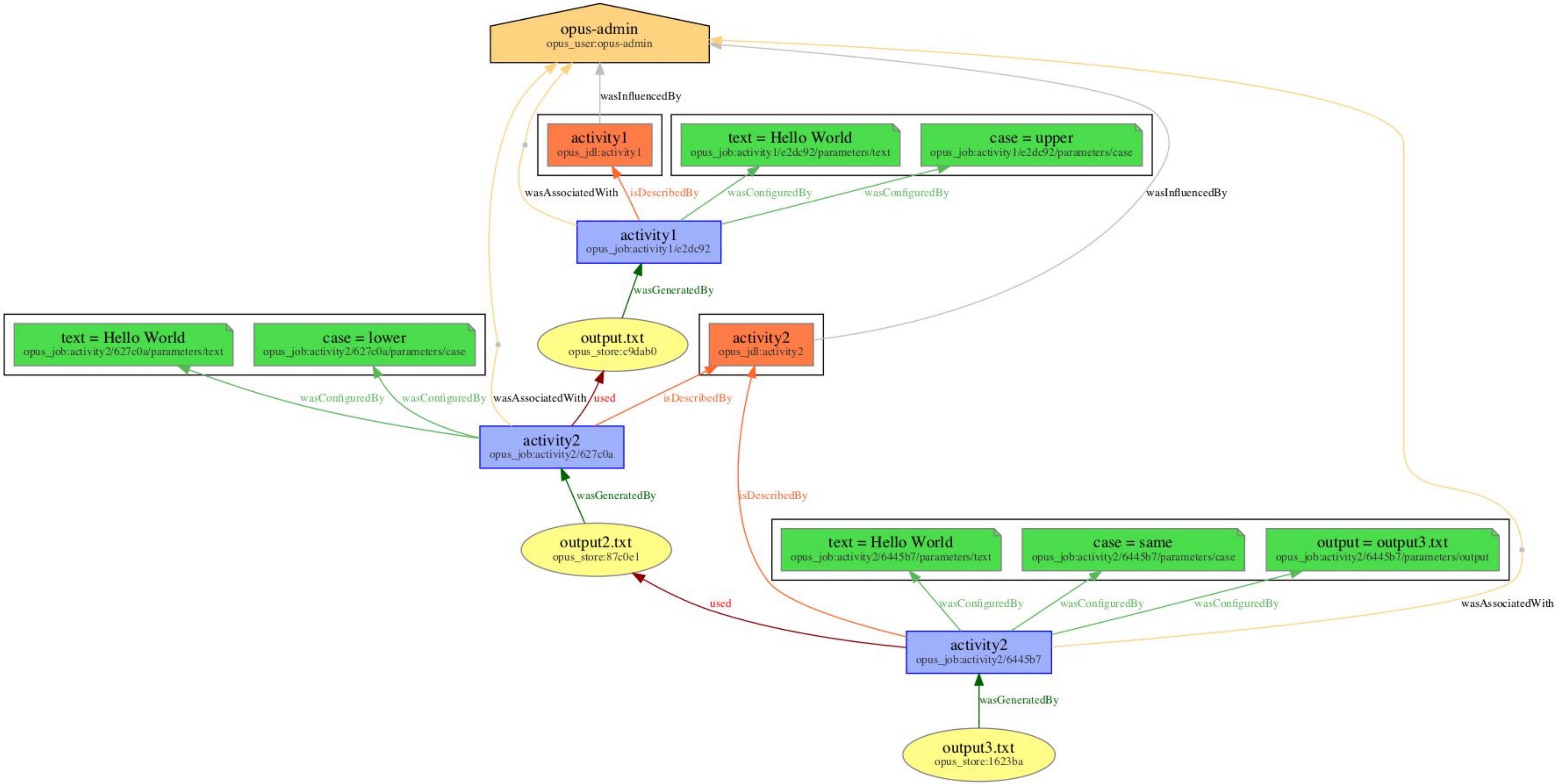
prov:label	case
prov:type	voprov:Parameter
prov:value	lower
voprov:annotation	case of output text in file
voprov:datatype	xs:string
voprov:default	same
voprov:options	same,lower,upper,camel
voprov:required	true

prov:label	activity2 description
prov:type	prov:Collection
prov:type	voprov:ActivityDescription
uws:executionDuration	20
uws:quote	10
voprov:description	Job that adds a line to a text file
voprov:description	test job 2
voprov:name	activity2
voprov:type	test
voprov:version	1

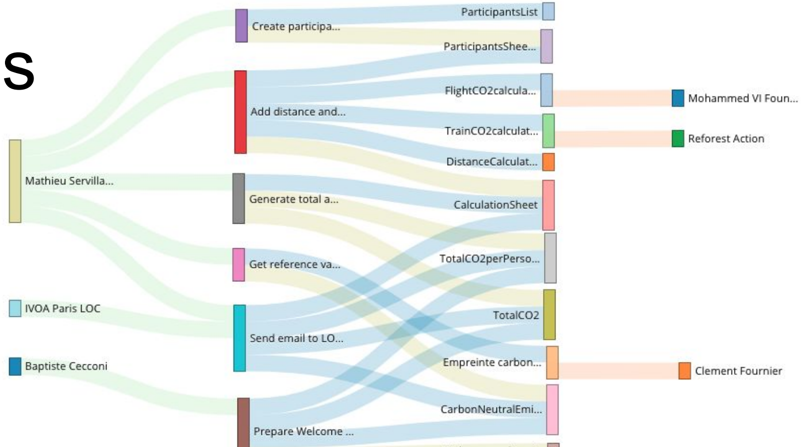
prov:startTime	2020-05-06T19:00:47
prov:endTime	2020-05-06T19:00:52
prov:label	activity2

+ ActivityDescription

+ Configuration


opus-admin
opus_user:opus-admin
wasInfluencedBy
activity1
opus_jdl:activity1
text = Hello World
opus_job:activity1/e2dc92/parameters/text
case = upper
opus_job:activity1/e2dc92/parameters/case
wasAssociatedWith
isDescribedBy
wasConfiguredBy
wasConfiguredBy
wasInfluencedBy
activity1
opus_job:activity1/e2dc92
wasGeneratedBy
text = Hello World
opus_job:activity2/627c0a/parameters/text
case = lower
opus_job:activity2/627c0a/parameters/case
output.txt
opus_store:c9dab0
activity2
opus_jdl:activity2
wasConfiguredBy
wasConfiguredBy
wasAssociatedWith
used
isDescribedBy
activity2
opus_job:activity2/627c0a
wasGeneratedBy
isDescribedBy
output2.txt
opus_store:87c0e1
text = Hello World
opus_job:activity2/6445b7/parameters/text
case = same
opus_job:activity2/6445b7/parameters/case
output = output3.txt
opus_job:activity2/6445b7/parameters/output
used
wasConfiguredBy
wasConfiguredBy
wasConfiguredBy
wasAssociatedWith
activity2
opus_job:activity2/6445b7
wasGeneratedBy
output3.txt
opus_store:1623ba

Carbon footprint of IVOA meetings

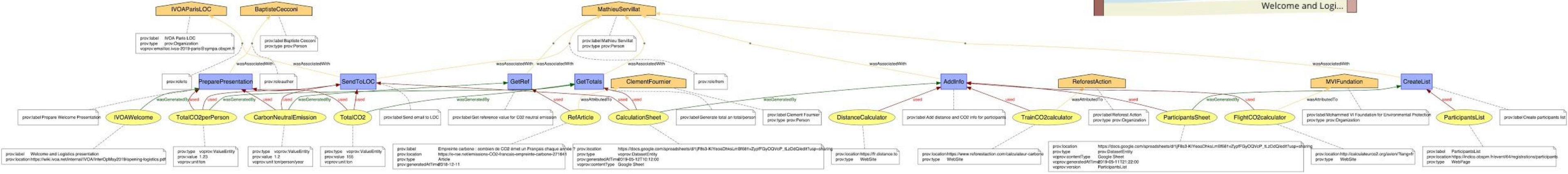


IVOA Paris 2019 - 132 participants

155 tons of CO2 for travels only

Check the detailed provenance of the calculations:

https://frama.link/CO2prov



IVOA Sydney 2020 virtual - 207 participants

300 grams of CO2 for visioconferences

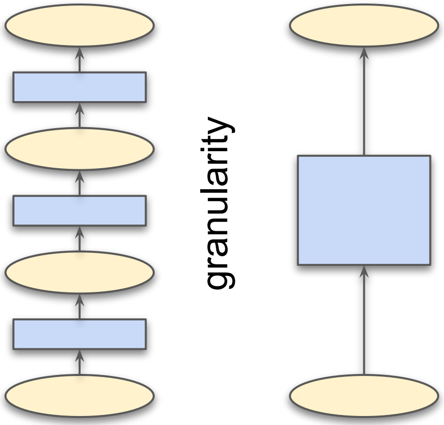
(~1500 h.person of visio estimated)

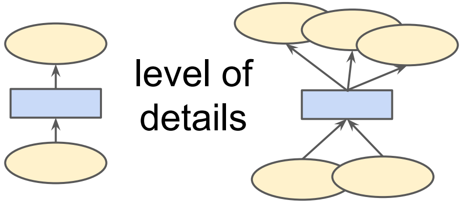
https://greenspector.com/en/which-video-conferencing-mobile-application-to-reduce-your-impact/

Applying the model

Different contexts in use cases

- Two flavours:
  - **on-top** (data products/collection already exist)
  - **inside** (save provenance information during the processing)
- **Identifiers:** unique and without meaning
- **Granularity** (what steps? what objects?)
- **Level of details** (descriptions? configuration?)





Different steps in provenance management

- How to **capture** the provenance information
- How to **store** this information
- How to **access** it
- How to **visualize** a provenance graph

Some terminology

- **full provenance**: graph/tree/chain of activities and entities up to the raw data. This information is not hosted by the entities themselves (stored on an external server? as separate files?)

- **minimum provenance**: attached to an entity, list of keywords that gives some context and info on **last activity** (general process/workflow, software versions, contacts...).
  Note: it would be interesting to include used entities, so that a full provenance may be reconstructed from each minimum provenance. However, such information on what was used may not be kept, or may not be complete.

- **end-user/specific "provenance"**: attached to an entity, 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, reco method...)
  *Note: may be extracted from full provenance (some parameters or entities generated at a given step), but it is considered as **data** here. Reversely, this specific "provenance" information may be a source of information to be mapped in the standard in order to fill the full provenance graph with more details.*

1. Defining the content of a **minimum provenance**
    - List of keywords related to the last activity and context
2. **Serializing** provenance
    - Both human and machine readable
    - Explore W3C formats, and YAML / VOTABLE / VOEvent formats
3. Provenance and **workflows**
    - Workflow information simply attached to provenance as used entities
    - Links with CWL, mapping
4. From provenance "**on-top**" to provenance "**inside**"
    - How to map specific provenance information into the IVOA model?
    - How to introduce provenance **capture** inside a pipeline?
5. Provenance **storage**
    - Database, interface and ingestion
6. Provenance **exploration** and **visualization**
    - Access protocols (ProvTAP, ProvSAP)
    - **`voprov`** Python package, or other tools