IVOA Provenance Data Model
Version 1.0

IVOA Working Draft 2018-05-30

Working group
DM

This version
http://www.ivoa.net/documents/ProvenanceDM/20180530

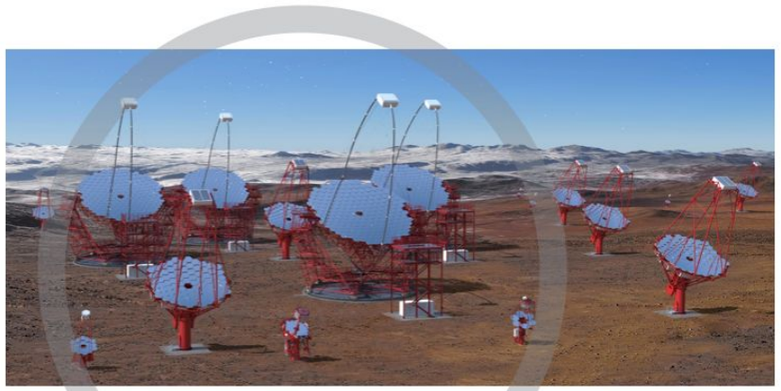
Latest version
http://www.ivoa.net/documents/ProvenanceDM

Previous versions
WD-ProvenanceDM-1.0-20170921.pdf
WD-ProvenanceDM-1.0-20161121.pdf
ProvDM-0.2-20160428.pdf
ProvDM-0.1-20141008.pdf

Author(s)
Kristin Riebe, Mathieu Servillat, François Bonnarel, Anastasia Galkin, Mireille Louys, Markus Nullmeier, Florian Rothmaier, Michèle Sanguillon, Ole Streicher, IVOA Data Model Working Group

Editor(s)
Kristin Riebe, Mathieu Servillat

Objectives and context



Observatory

Consortium

Data acquisition
(triggers, cameras, …)

Data Centers
(data processing, reduction, reconstruction, simulations, …)

Additional data

End User

Archive, Dissemination

Web portal

Publications

- Data product generation **obscure** to end user
- **Quality**, **reliability**, **trustworthiness**?
- **Usefulness** of the data?

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

Use cases

- CTA pipeline and data access
- RAVE (Radial Velocity Experiment)
- POLLUX (synthetic stellar spectra service)
- SVOM gamma ray burst / transients
- TAP-based API for images in an archive @CDS
- APPLAUSE database
- MUSE WISE pipeline

⇒ Different aspects of Provenance

- How to **collect** the provenance information
- How to **store** this information
- How to **access** and **visualize** the provenance

Draft content

1 Introduction

- 1.1 Goal of the provenance model
- 1.2 Minimum requirements for provenance
- 1.3 Role within the VO architecture
- 1.4 Previous efforts

2 The provenance data model

- 2.1 Overview: Conceptional UML class diagram and introduction to core classes
- 2.2 Model description
  - 2.2.1 Class diagram and VO-DML compatibility
  - 2.2.2 Entity (and EntityDescription)
  - 2.2.3 Collection
  - 2.2.4 Activity and ActivityDescription
  - 2.2.5 ActivityFlow
  - 2.2.6 Entity-Activity relations
  - 2.2.7 Parameters
  - 2.2.8 Agent

define sub-entities

separate section

link to ProvONE?

separate section

3 Serialization of the provenance data model

- 3.1 Introduction
- 3.2 Serialization formats: PROV-N, PROV-JSON and PROV-XML
- 3.3 PROV-VOTable format
- 3.4 Serialization of description classes for web services
- 3.5 W3C PROV-DM compatible serializations

comes with ProvTAP

can all IVOA concepts fit in W3C serializations?

4 Accessing provenance information

Appendix A Serialization Examples

Appendix B Links to other data models

- B.1 Links with Dataset/ObsCore Model
- B.2 Links with Simulation Data Model

Goals

A: Tracking the production history

Find out which steps were taken to produce a dataset and list the methods/tools/software that was involved.

B: Attribution and contact information

Find the people involved in the production of a dataset, that need to be cited or can be asked for more information.

C: Locate error sources

Find the location of possible error sources in the generation of a dataset.

D: Quality assessment

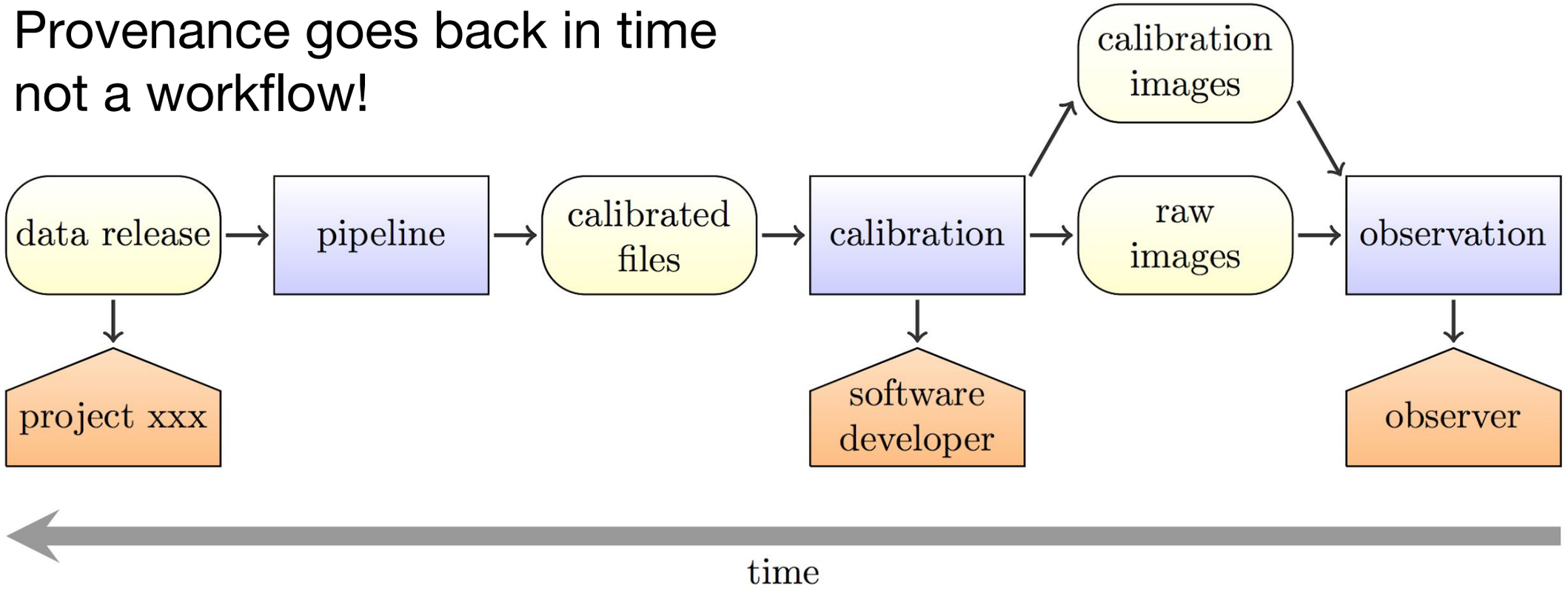
Judge the quality of an observation, production step or dataset.

E: Search in structured provenance metadata

This would allow one to also do a “forward search”, i.e. locate derived datasets or outputs.

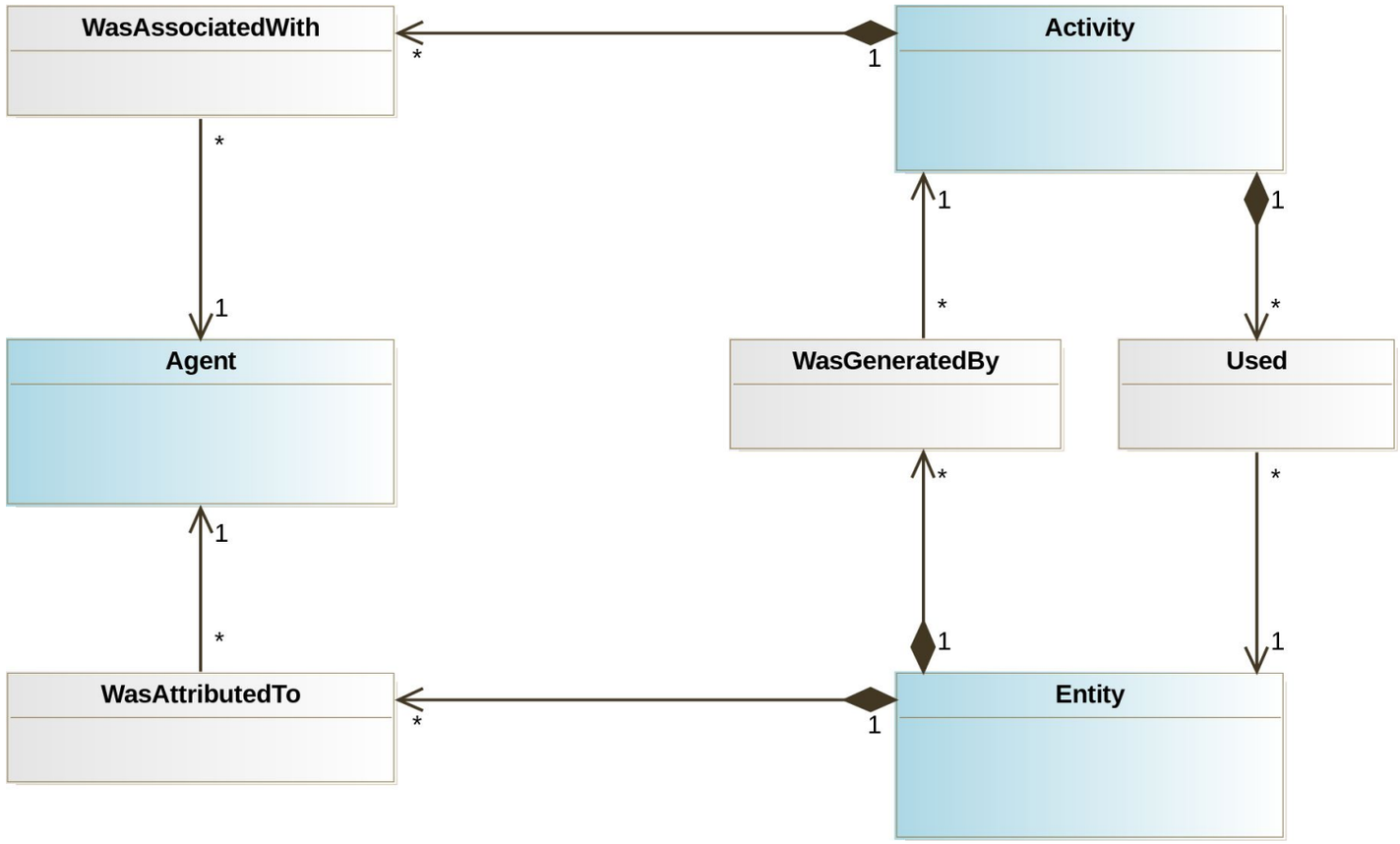
What is provenance?

Provenance goes back in time
not a workflow!



- **Provenance** = Identify how a data product was produced
- **Configuration** = Identify what detailed options were used
- **Contextual** information:
  - Instrument Configuration
  - Ambient Conditions
  - Software environment

Core Provenance Data Model



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

30 April 2013

Word Wide Web Consortium

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

- Core concepts from the W3C PROV recommendations
  - **Entity - Activity - Agent**
  - **Relations** and **roles** = provenance information
  - W3C PROV has many more relations
  - IVOA Provenance connected to **VO concepts** and **astronomy needs**

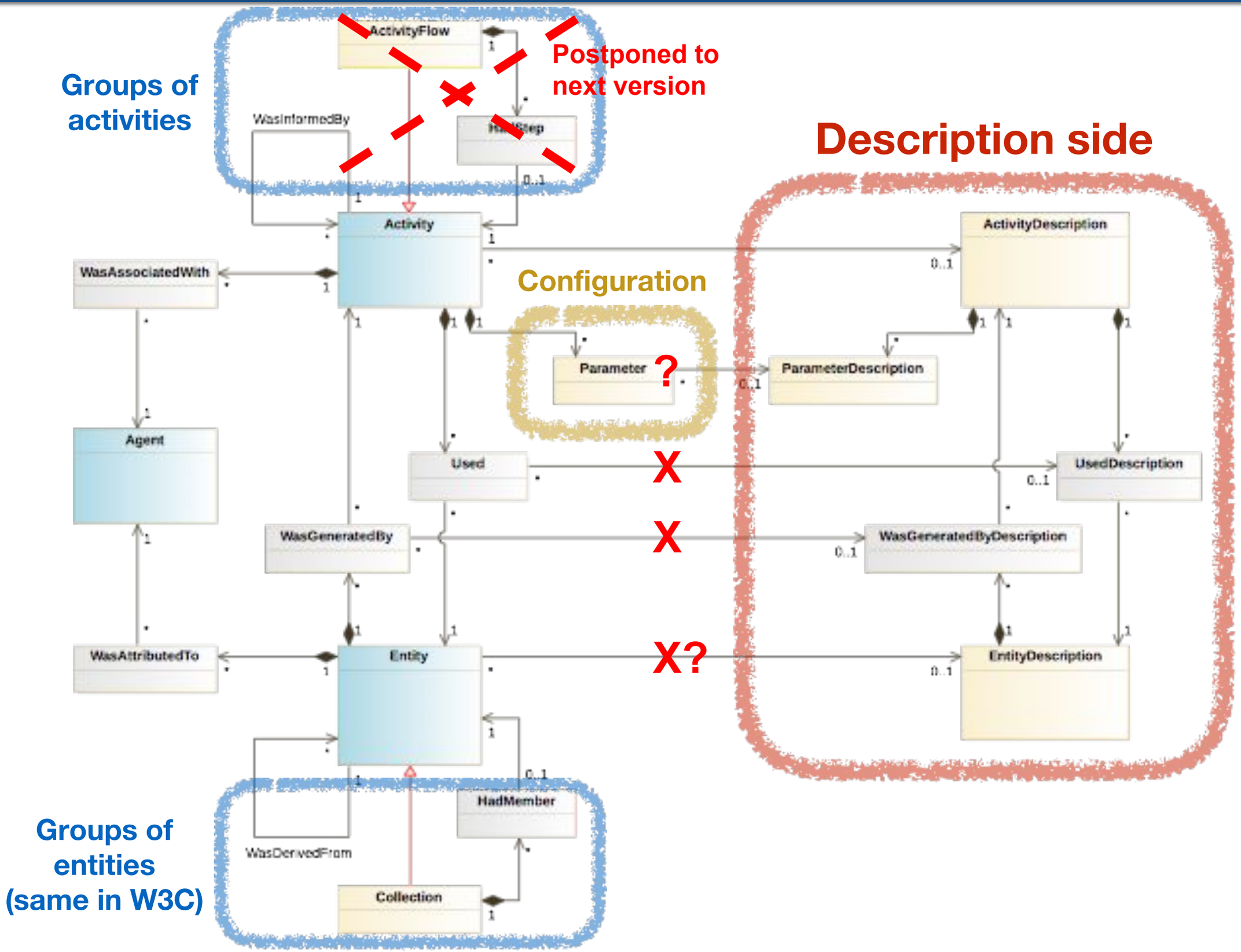
Concepts In Astronomy

- **Entities**: datasets composed of VOTables, FITS files or database tables, or files containing logs, values (spectra, lightcurves), parameters, etc.
- **Activities**: an observation, a simulation, or processing steps (image stacking, object extraction, etc.).
- **Agents**: the people involved can be individual persons (observer, publisher...), groups or organisations.
- **Connections to existing VO concepts**
  - Entity <—> Dataset (Curation, DataID), ObsCore, SimDM DataObject
  - Activity <—> SimDM (Resource, Experiment)
  - Agent <—> Party, Contact
- **Connections to external concepts** (PROV, ProvONE, DOI, ORCID, …)

Minimum requirements

1. Provenance information must be stored in a **standard model**, with **standard serialization formats**.
2. Provenance information must be **machine readable**.
3. Provenance data model classes and attributes should be **linked to other IVOA concepts** when relevant (DatasetDM, ObsCoreDM, SimDM, VOTable, UCDs...).
4. Provenance information should be **serializable into the W3C provenance standard formats** (PROV-N, PROV-XML, PROV-JSON) with minimum information loss.
5. Provenance metadata must contain information to find immediate **progenitor(s)** (if existing) for a given entity, i.e. a dataset.
6. An entity must point to **the activity that generated it** (if the activity is recorded).
7. Activities must point to **input entities** (if applicable).
8. Activities may point to **output entities**.
9. Provenance information should make it possible to derive the **chronological** sequence of activities.
10. Provenance information can only be given for **uniquely identifiable entities**, at least inside their domain.
11. Released entities should have a **main contact**.
12. It is recommended that all activities and entities have **contact information** and contain a (short) **description** or link to a description.

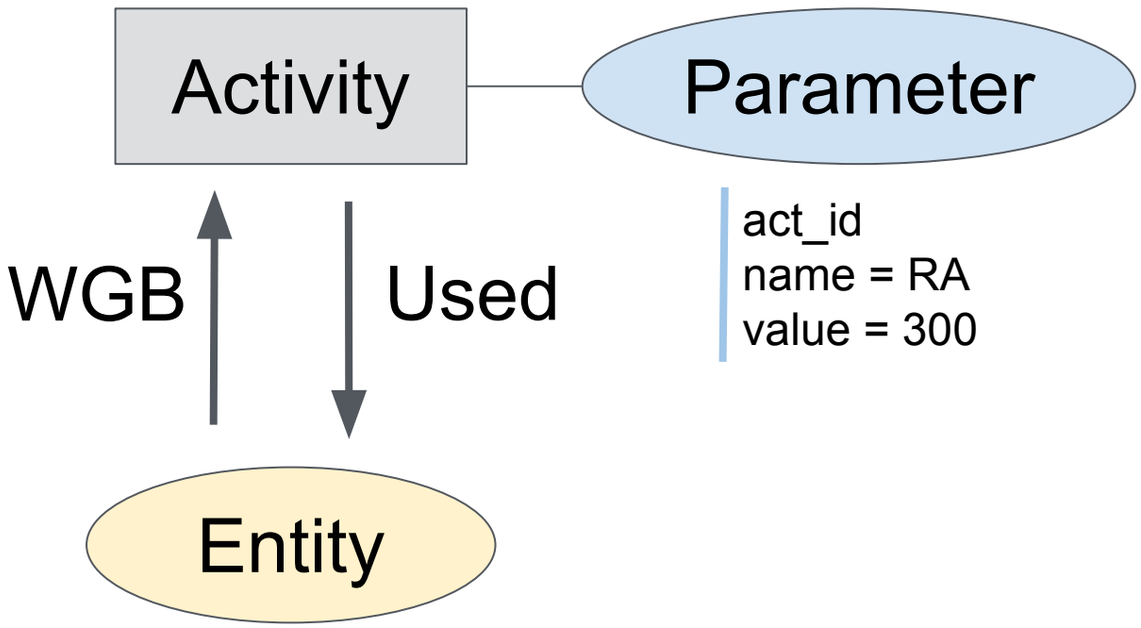
IVOA Provenance Data Model diagram



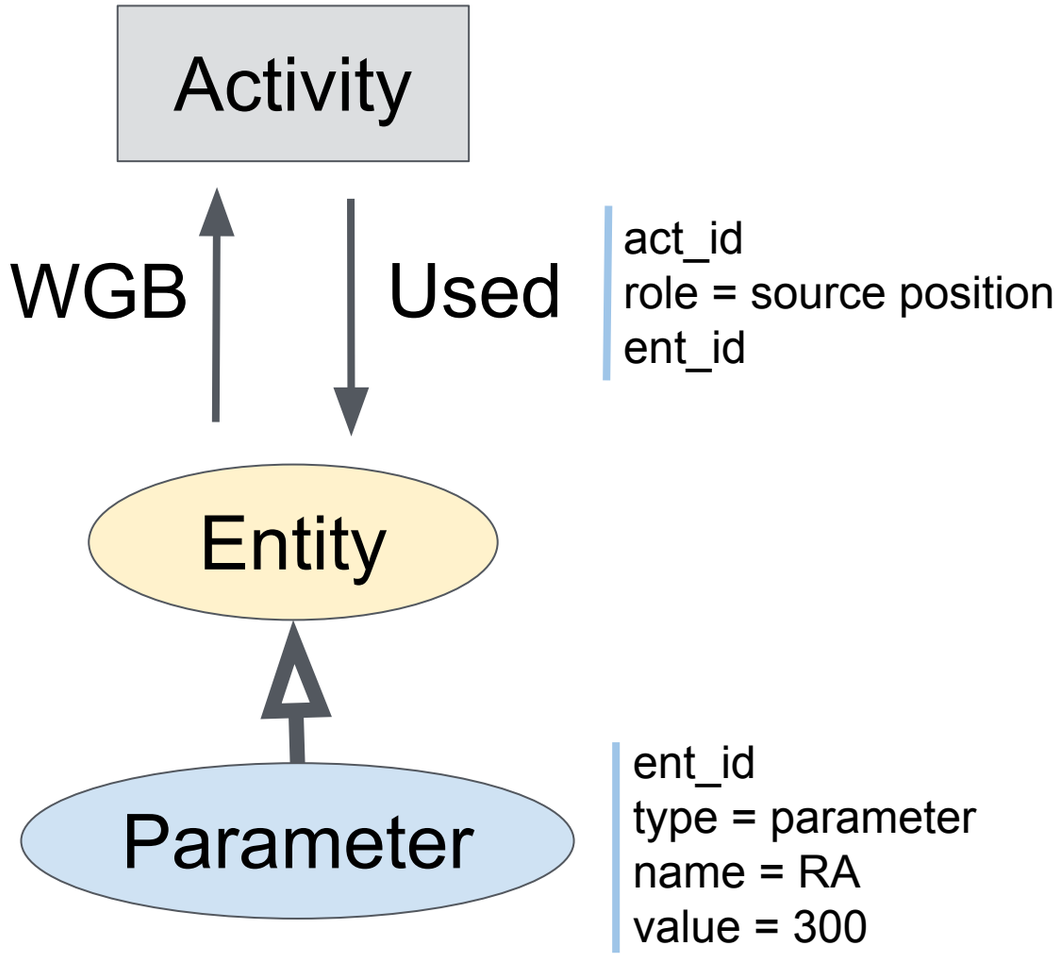
Parameter (section 2.2.7)

name = RA
unit = deg
ucd = pos.eq.ra
utype = ...
description = source position

name = RA
unit = deg
ucd = pos.eq.ra
utype = ...



act_id
name = RA
value = 300



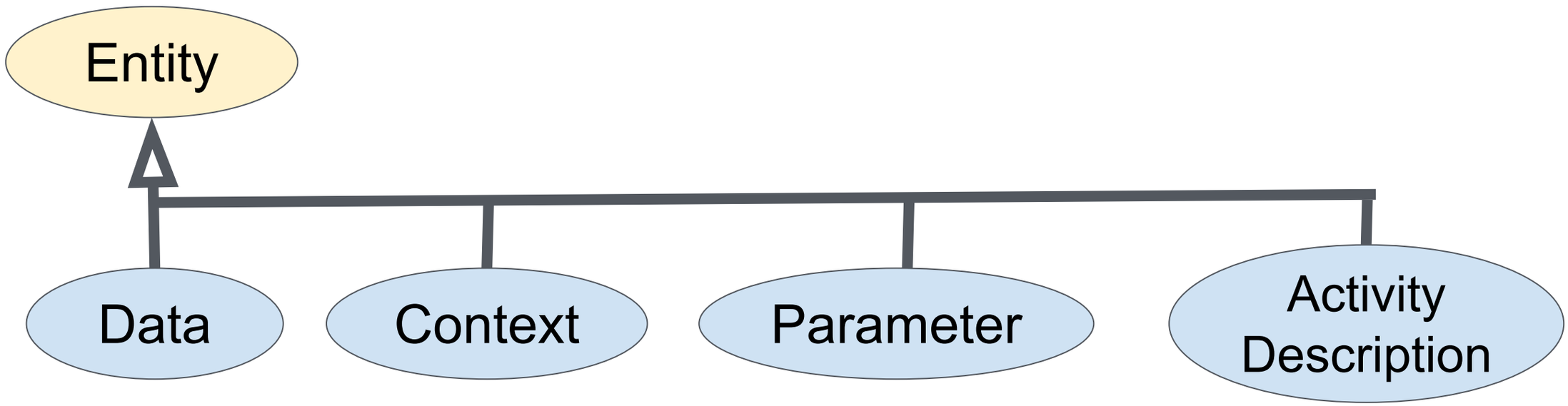
act_id
role = source position
ent_id

ent_id
type = parameter
name = RA
value = 300

- Minimal representation
- logical place for configuration

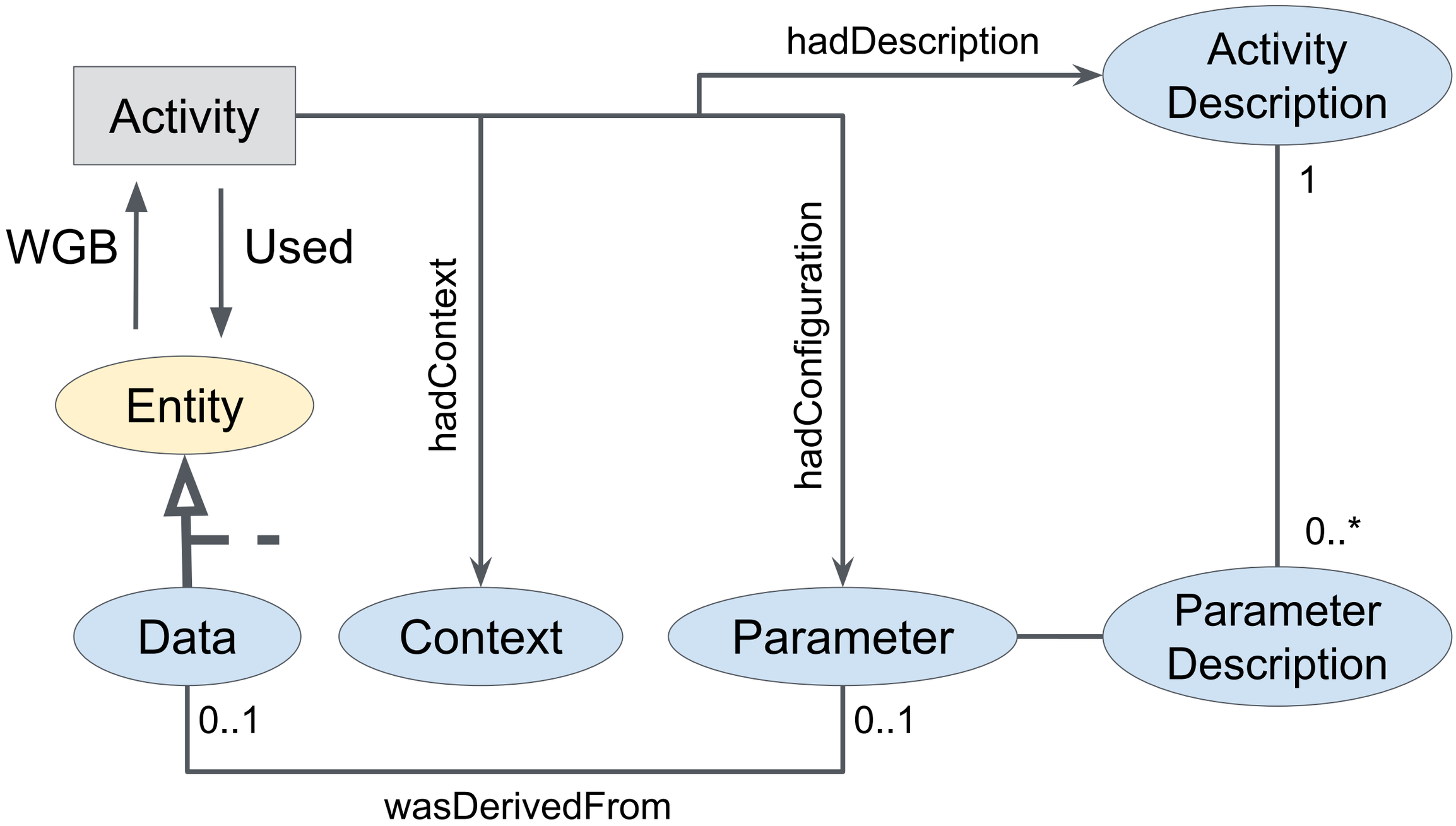
- Reusable (e.g. as a pointing position)
- Provenance of the parameter

Specialization of entities



- **Entity**: An entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary. (W3C PROV definition)
- **Data**: digital, machine-readable information in some content that will be used/transformed/analyzed... Could be a cell or a column in a table, a file, an image, a cube... The **value** attribute could directly transport this information, or the **location** attribute directs to it.
- **Document**: information presented in a human readable form.
- **Thing**: a physical object, such as a tool, an instrument, a telescope...
- **ActivityDescription/Plan**: explanations on the activity, main contact(s), inputs expected, output produced...
- **ConfigFile:** file containing configuration information for the activity.
- **Parameter:** configuration of the activity before execution as a key=value parameter in the IVOA framework.
- **AmbientConditions**: common, prevailing, and uncontrolled atmospheric and weather conditions in a room or place that influence the activity
- **InstrumentConfiguration:** description of how the structure of the instrument and how it is set up before an activity starts.
- **ExecutionEnvironment**: describes a particular execution platform, such as an operating system or a database management system. Execution environments are used to describe the context in which the execution of an activity takes place. Execution environments could also describe the computing hardware of a system.

Specialized relations



Serializations - W3C PROV formats

```
<prov:document xmlns:ctadata="ivo://vopdc.obspm/cta#" xmlns:ctajob
  <prov:activity prov:id="ctajobs:ctbin">
    <prov:startTime> 2016-03-13T23:44:46 </prov:startTime>
    <prov:endTime> 2016-03-13T23:44:56 </prov:endTime>
  </prov:activity>
  <prov:agent prov:id="cta:consortium">
    <prov:type xsi:type="xsd:string"> Organization </prov:type>
  </prov:agent>
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="ctajobs:ctbin" />
    <prov:agent prov:ref="cta:consortium" />
  </prov:wasAssociatedWith>
  <prov:entity prov:id="uwsdata:parameters/inobs" />
  <prov:used>
    <prov:activity prov:ref="ctajobs:ctbin" />
    <prov:entity prov:ref="uwsdata:parameters/inobs" />
  </prov:used>
  <prov:entity prov:id="uwsdata:results/outcube" />
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="uwsdata:results/outcube" />
    <prov:activity prov:ref="ctajobs:ctbin" />
  </prov:wasGeneratedBy>
  <prov:wasDerivedFrom>
    <prov:generatedEntity prov:ref="uwadata:results/outcube" />
    <prov:usedEntity prov:ref="uwsdata:parameters/inobs" />
  </prov:wasDerivedFrom>
  <prov:entity prov:id="uwsdata:results/logfile" />
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="uwsdata:results/logfile" />
    <prov:activity prov:ref="ctajobs:ctbin" />
  </prov:wasGeneratedBy>
  <prov:wasDerivedFrom>
    <prov:generatedEntity prov:ref="uwsdata:results/logfile" />
    <prov:usedEntity prov:ref="uwsdata:parameters/inobs" />
  </prov:wasDerivedFrom>
</prov:document>
```

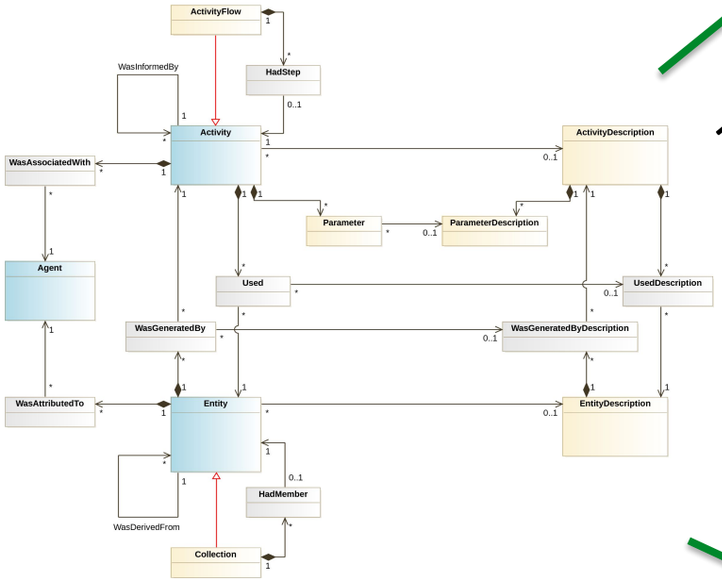
```
{
  - wasAssociatedWith: {
      - _:id1: {
            prov:agent: "cta:consortium",
            prov:activity: "cta:anactools_v1.1"
        }
    },
  - agent: {
      - cta:consortium: {
            prov:type: "Organization"
        }
    },
  - entity: {
        uwsdata:results/fit_results: { },
        uwsdata:results/configfile: { },
        uwsdata:results/butterfly: { },
        uwsdata:results/spectrum_plot: { },
        uwsdata:results/spectrum: { }
    },
  - prefix: {
        uwsdata: "https://voparis-uws-test.obspm.fr/rest
        cta: "http://www.cta-observatory.org#",
        voprov: "http://www.ivoa.net/ns/voprov#"
    },
  - activity: {
      - cta:anactools_v1.1: {
            prov:startTime: "2016-04-07T00:26:00",
            prov:endTime: "2016-04-07T00:27:15"
        }
    },
  - wasGeneratedBy: {
      - _:id5: {
            prov:entity: "uwsdata:results/butterfly",
            prov:activity: "cta:anactools_v1.1"
        },
      - _:id4: {
            prov:entity: "uwsdata:results/fit_results",
            prov:activity: "cta:anactools_v1.1"
        },
```

Serializations - VOTable

```
<?xml version="1.0" encoding="UTF-8"?>
<VOTABLE version="1.2" xmlns="http://www.ivoa.net/xml/VOTable/v1.2"
    xmlns:ex="http://www.example.com/provenance"
    xmlns:ivo="http://www.ivoa.net/documents/rer/ivo/"
    xmlns:voprov="http://www.ivoa.net/documents/dm/provdm/voprov/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.ivoa.net/xml/VOTable/v1.2 http://www.ivoa.net/xml/VOTable/VOTable-1.2.xsd">
<RESOURCE type="provenance">
    <DESCRIPTION>Provenance VOTable</DESCRIPTION>
    <TABLE name="Usage" utype="voprov:used">
        <FIELD arraysize="*" datatype="char" name="activity" ucd="meta.id" utype="voprov:Usage.activity"/>
        <FIELD arraysize="*" datatype="char" name="entity" ucd="meta.id" utype="voprov:Usage.entity"/>
        <DATA>
            <TABLEDATA>
                <TR>
                    <TD>ex:Process1</TD>
                    <TD>ivo://example#DSS2.143</TD>
                </TR>
            </TABLEDATA>
        </DATA>
    </TABLE>
    <TABLE name="Generation" utype="voprov:wasGeneratedBy">
        <FIELD arraysize="*" datatype="char" name="entity" ucd="meta.id" utype="voprov:Generation.entity"/>
        <FIELD arraysize="*" datatype="char" name="activity" ucd="meta.id" utype="voprov:Generation.activity"/>
        <DATA>
            <TABLEDATA>
                <TR>
                    <TD>ivo://example#Public_NGC6946</TD>
                    <TD>ex:Process1</TD>
                </TR>
            </TABLEDATA>
        </DATA>
    </TABLE>
    <TABLE name="Activity" utype="voprov:Activity">
        <FIELD arraysize="*" datatype="char" name="id" ucd="meta.id" utype="voprov:Activity.id"/>
        <FIELD arraysize="*" datatype="char" name="name" ucd="meta.title" utype="voprov:Activity.name"/>
        <FIELD arraysize="*" datatype="char" name="start" ucd="" utype="voprov:Activity.startTime"/>
        <FIELD arraysize="*" datatype="char" name="stop" ucd="" utype="voprov:Activity.endTime"/>
        <DATA>
            <TABLEDATA>
                <TR>
                    <TD>ex:Process1</TD>
```

Serialization context

IVOA DM



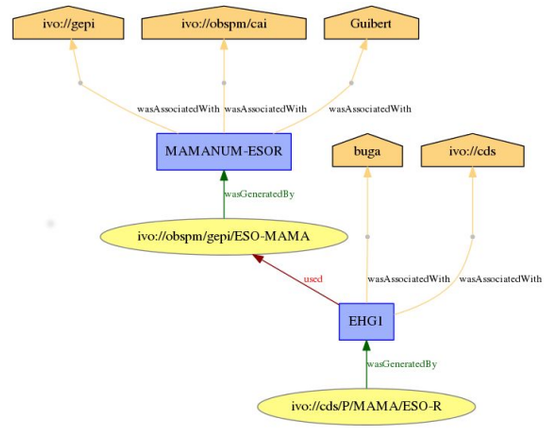
All classes defined in the specification

Translation

W3C PROV View

PROV-JSON
PROV-N
PROV-XML

ProvSAP



W3C Tools
Provenance suite,
Prov-Python

Translation

ProvTAP

IVOA View

VOTable
JSON
FITS

TopCat

TapHandle

Serializations - ActivityDescription

```
<RESOURCE ID="gammapy_maps" name="gammapy_maps" type="meta" utype="voprov:ActivityDescription">

    <DESCRIPTION>Use gammapy to generate a count map from a list of observations</DESCRIPTION>
    <!-- Service Descriptor -->
    <PARAM name="accessURL" datatype="char" arraysize="*" value="https://voparis-uws-test/rest/gammapy_maps" />
    <PARAM name="standardID" datatype="char" arraysize="*" value="ivo://ivoa.net/std/SODA#1.0" />
    <!-- Activity Description -->
    <PARAM name="type" datatype="char" arraysize="*" value="None" utype="voprov:ActivityDescription.type"/>
    <PARAM name="subtype" datatype="char" arraysize="*" value="None" utype="voprov:ActivityDescription.subtype"/>
    <PARAM name="annotation" datatype="char" arraysize="*" value="Use gammapy to generate a count map from a list of
    <PARAM name="version" datatype="char" arraysize="*" value="None" utype="voprov:ActivityDescription.version"/>
    <PARAM name="doculink" datatype="char" arraysize="*" value="https://luthgitlab.obspm.fr/jlefaucheur/hess_release
    <PARAM name="contact_name" datatype="char" arraysize="*" value="Julien Lefaucheur" utype="voprov:Agent.name"/>
    <PARAM name="contact_email" datatype="char" arraysize="*" value="" utype="voprov:Agent.email"/>
    <!-- UWS job attributes -->
    <PARAM name="executionDuration" datatype="int" value="600" utype="uws:Job.executionDuration"/
    <PARAM name="quote" datatype="int" value="120" utype="uws:Job.quote"/>

    <!-- UWS parameters (Provenance Entities or Parameters) -->
    <GROUP name="InputParams">
        <PARAM ID="obs_ids" arraysize="*" datatype="char" name="obs_ids" value="47802 47803 47804
            <DESCRIPTION>List of runs</DESCRIPTION>
        </PARAM>
        <PARAM ID="RA" datatype="double" name="RA" value="329.7169379" unit="deg"...>
        <PARAM ID="Dec" datatype="double"
        <PARAM ID="nxpix" arraysize="*" da
            <DESCRIPTION>Number of pixels
            <VALUES>
                <MIN value="0"/>
                <MAX value="1000"/>
            </VALUES>
        </PARAM>
        <PARAM ID="nypix" arraysize="*" da
        <PARAM ID="binsz" datatype="float"
    </GROUP>
```

VOTable
DataLink Service Descriptor
UWS Job Description Language
Provenance ActivityDescription

```
    <!-- Used Entities -->
    <GROUP name="Used">
        <GROUP name="obs_ids" utype="voprov:UsedDescription" ref="obs_ids">
            <PARAM arraysize="*" datatype="char" name="role" utype="voprov:UsedDescription.role" value="DL3"/>
            <PARAM arraysize="*" datatype="char" name="location" utype="voprov:EntityDescription.location" value=""/
            <PARAM arraysize="*" datatype="char" name="content_type" utype="voprov:EntityDescription.content_type" v
        </GROUP>
    </GROUP>

    <!-- Generated Entities / UWS results -->
    <GROUP name="Generated" utype="voprov:WasGeneratedBy">
        <GROUP name="count_map" utype="voprov:EntityDescription">
            <DESCRIPTION>Count map</DESCRIPTION>
            <PARAM arraysize="*" datatype="char" name="role" utype="voprov:UsedDescription.role" value="DL4 image"/>
            <PARAM arraysize="*" datatype="char" name="default" utype="voprov:Entity.id" value="count_map.fits"/>
            <PARAM arraysize="*" datatype="char" name="content_type" utype="voprov:EntityDescription.content_type" v
        </GROUP>
        <GROUP name="count_preview" utype="voprov:EntityDescription">
            <DESCRIPTION>Count map preview</DESCRIPTION>
```

Conclusions

- Many sections in the draft are stable
- ProvSAP and ProvTAP moved to DAL drafts
- Implementation note based on many use cases
- Still some open questions:
  - modeling of Parameter
  - relations with Description classes
  - mapping for valid W3C serialization without loss
- Next steps
  - move to RFC track before next IVOA meeting