IVOA Provenance Data Model

Version 1.0

IVOA Proposed Recommendation 2018-10-15

Working group
DM

This version
http://www.ivoa.net/documents/ProvenanceDM/20181015

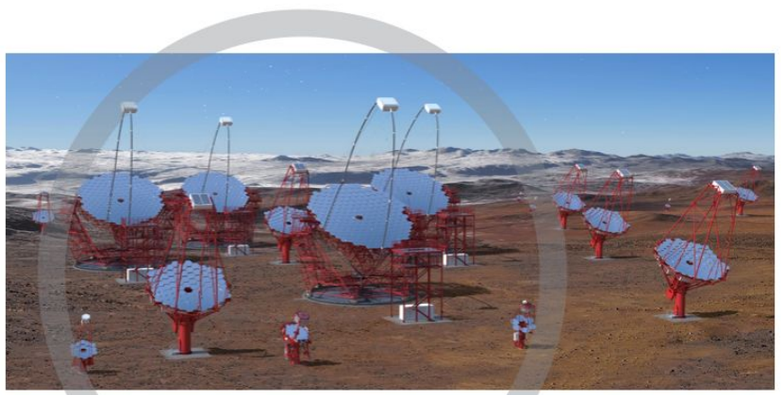
Latest version
http://www.ivoa.net/documents/ProvenanceDM

Previous versions
WD-ProvenanceDM-1.0-20180530.pdf
WD-ProvenanceDM-1.0-20170921.pdf
WD-ProvenanceDM-1.0-20161121.pdf
ProvDM-0.2-20160428.pdf
ProvDM-0.1-20141008.pdf

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

Editor(s)
Mathieu Servillat

Objectives and context



Observatory

Consortium

Data acquisition
(triggers, cameras, …)

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

Archive, Dissemination

Web portal

Additional data

End User

Publications

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

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

Definition of Provenance

WIKIPEDIA

Provenance (from the French provenir, 'to come from/forth') is the **chronology of the ownership**, **custody** or **location** of a historical object. [...]

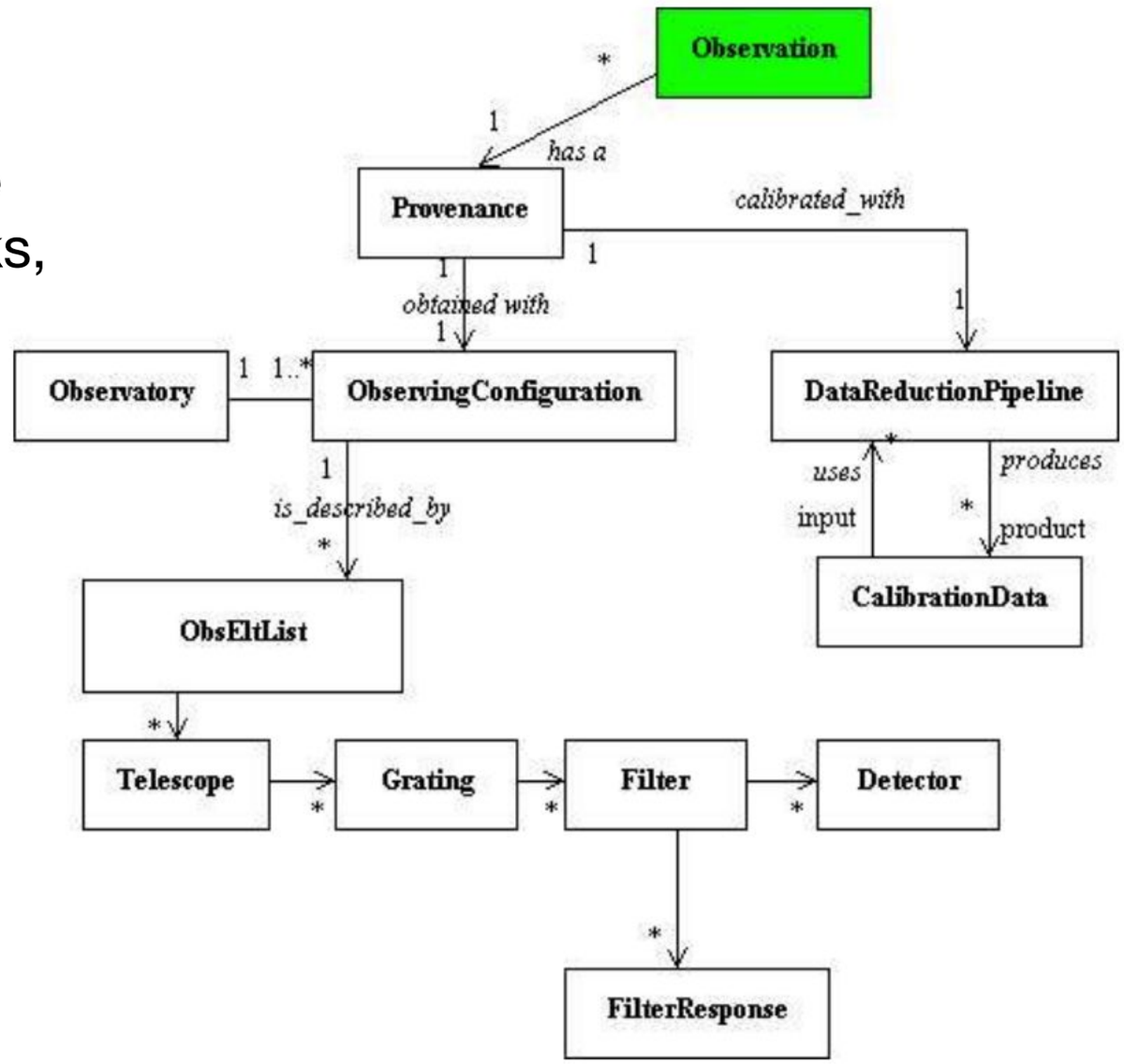
The primary purpose of tracing the provenance of an object or entity is normally to **provide contextual** and **circumstantial evidence** for its **original production** or discovery, by establishing, as far as practicable, its later history, especially the sequences of its formal **ownership**, **custody** and **places of storage**. The practice has a particular value in helping **authenticate** objects. [...] establishing provenance is essentially a **matter of documentation**.

Early definition of IVOA Provenance

Observation DM, 2005

Jonathan McDowell

The Provenance is the description of **how the dataset was created.** For many analysis tasks, information about **some aspect** of the data acquisition chain is needed.



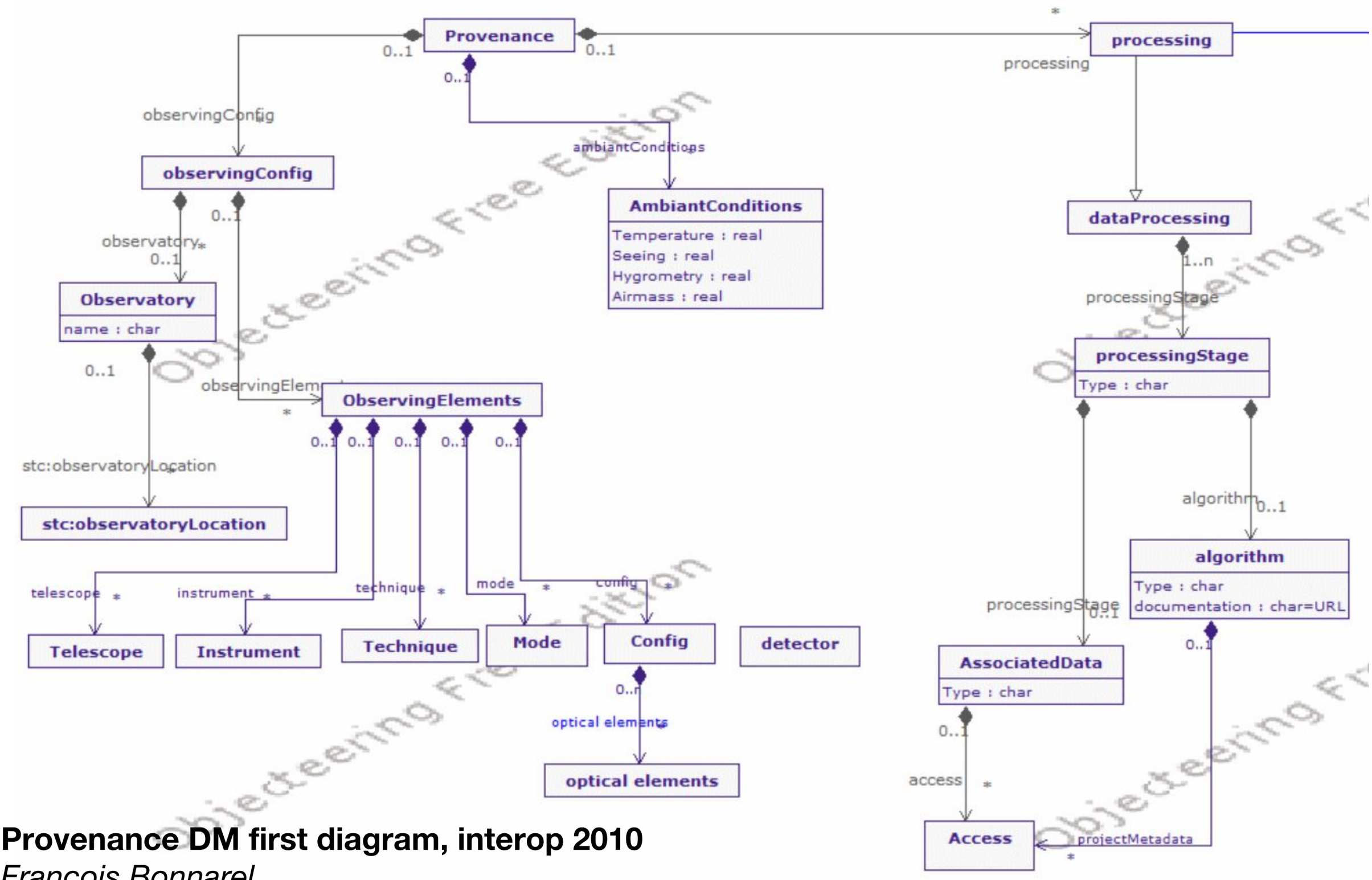
Provenance DM first diagram, interop 2010 **(next slide)**

François Bonnarel

Motivation: **trace the history** of the dataset

- Some kind of **quality** assessment
  - What were the **ambient conditions**?
  - How was the telescope **configured**?
  - What kind of **processing** has been done?
- Describe the previous steps and access to progenitors if reprocessing of the data is needed

Early definition of IVOA Provenance



Provenance DM first diagram, interop 2010

François Bonnarel

W3C Provenance definition

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

W3C PROV (PROV-DM, 2013)

Provenance is defined as a **record that describes the people, institutions, entities, and activities** involved in **producing**, **influencing**, or **delivering** a **piece of data or a thing**.

In particular, the provenance of information is crucial in deciding whether information is to be **trusted**, how it should be **integrated** with other diverse information sources, and how to **give credit** to its originators when reusing it.

Core concepts from the W3C PROV recommendations:

- **Entity - Activity - Agent**
- **Relations** and **roles**: e.g. generation, usage, influence, association, attribution, derivation, information
- W3C PROV has more relations (see components and extensions)
- IVOA Provenance connected to **VO concepts** and **astronomy needs**

Use cases

- CTA (Cherenkov Telescope Array) data processing and access
- RAVE (Radial Velocity Experiment)
- POLLUX (synthetic stellar spectra service)
- CDS image databases
- SVOM gamma ray burst / transients
- APPLAUSE photographic plates database
- MuseWise pipeline

⇒ Different aspects of Provenance

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

Implementations

- Cherenkov Telescope Array preparation
  - Python Provenance class dedicated to the **capture** of provenance information
  - OPUS: a provenance enhanced job controller, **storing** and **exporting** provenance information for jobs → **Apps2 talk (today 4pm)**
- Pollux, RAVE, MuseWise
  - Different implementations of a simple **access** protocol to the provenance metadata → **DAL2 talk (Saturday 11am)**
- CDS HiPS database
  - Implementation of a **TAP** service for Prov metadata bound to ObsTAP metadata → **ADASS talk O11-3 (Monday 4:30m)**
- CDS test image database
  - A **triple store** implementation for an Aladin image database → **ADASS poster P11-5**
- SVOM Quick Analysis
  - Track and use provenance information to **reprocess** the raw data with refined parameters

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.

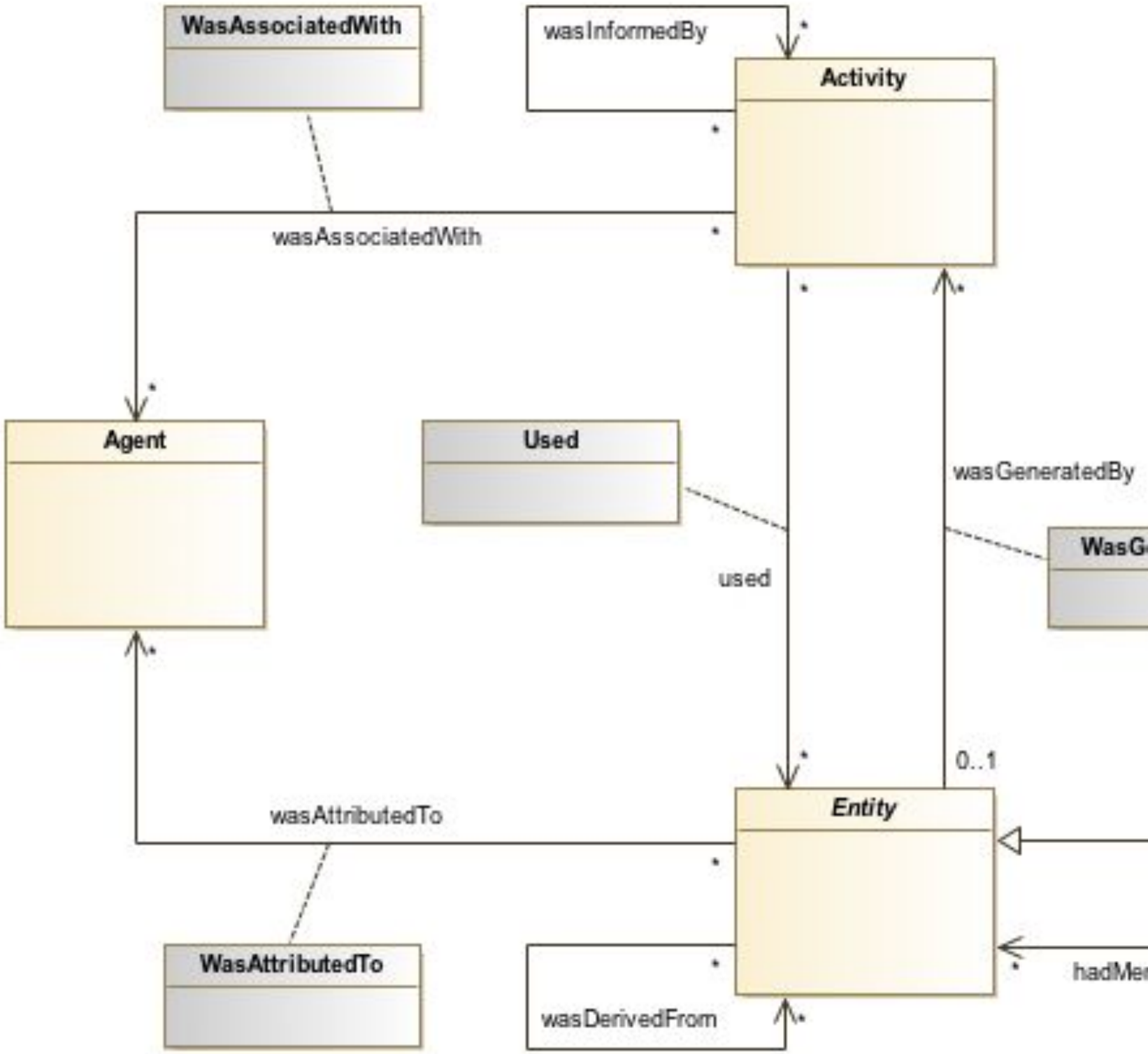
E: Search in structured provenance metadata

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

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 IVOA semantics, data models and formats** (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 be linked to **the activity that generated it** (if the activity is recorded).
7. Activities must be linked 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. Entities, Activities and Agents must be **uniquely identifiable** within a 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.

Core Provenance Data Model



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

W3C

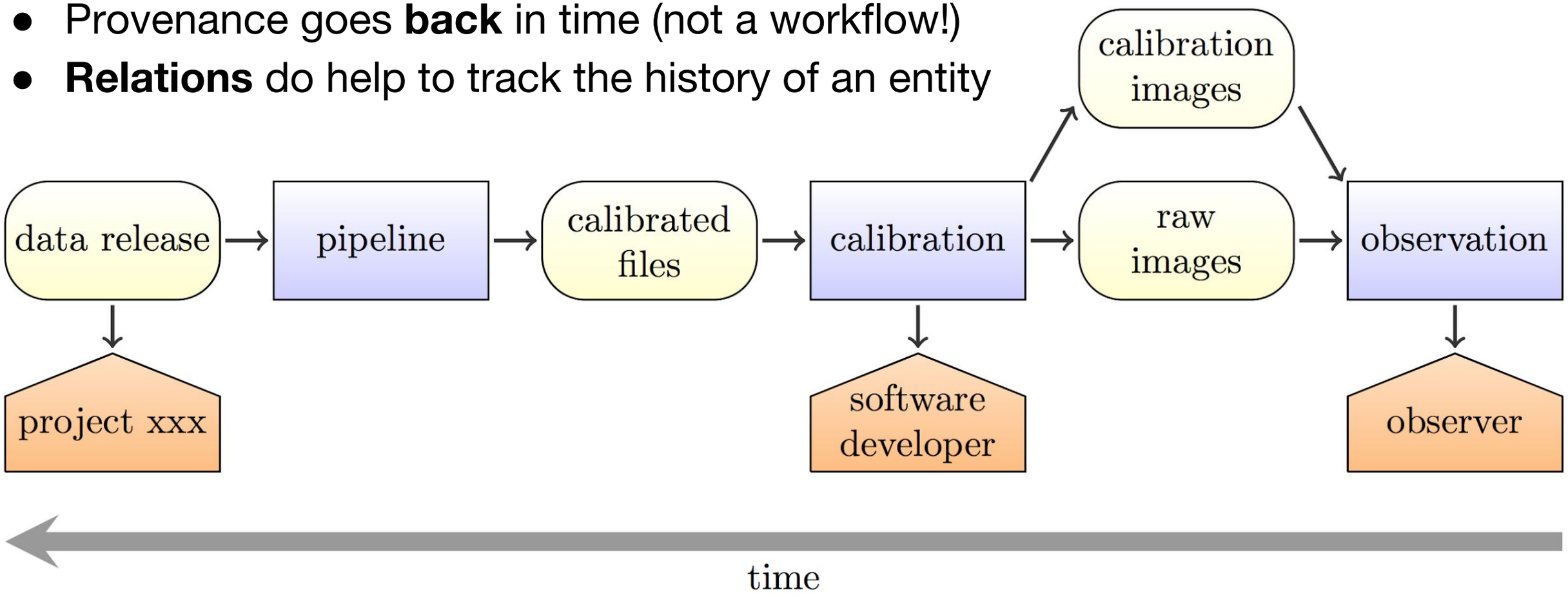
Word Wide Web Consortium

Existing Implementations:

- Java library
- Python package
- ProvStore web service

Chain of entities, activities, agents

- Provenance goes **back** in time (not a workflow!)
- **Relations** do help to track the history of an entity



- Does this answers all our goals? A, B: **yes**, but C, D, E: **no**
- The relevant provenance **information** that will help assess the quality (D), locate errors (C) and enable searches in provenance metadata (E) should be contained by **additional entities**
- The core model is **too generic** to provide this information

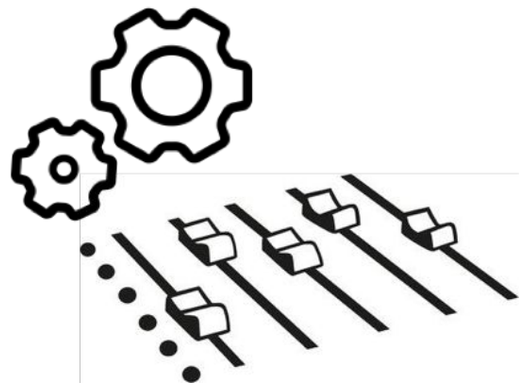
Relevant provenance information?

How was the calibration performed, which steps, which algorithms?

Description: information on the expected working of an activity and on the expected structure of an entity. This descriptive information is what is known before any activity or entity instance is created,

What was the detailed configuration of the pipeline?

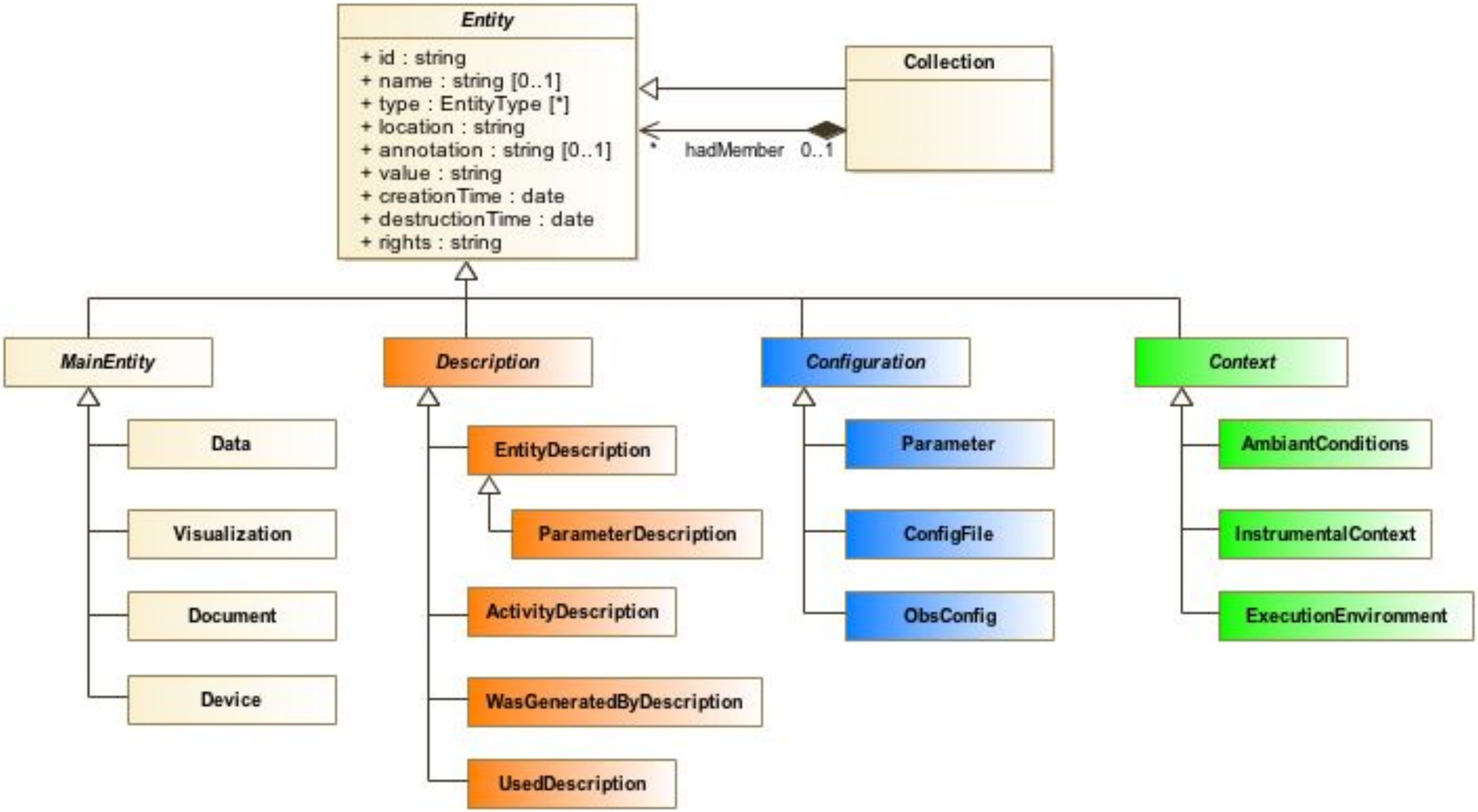
Configuration: information passed to an activity in order to configure its execution and which directly influences the development of the activity (e.g. Parameter, Config File, ObsConfig),



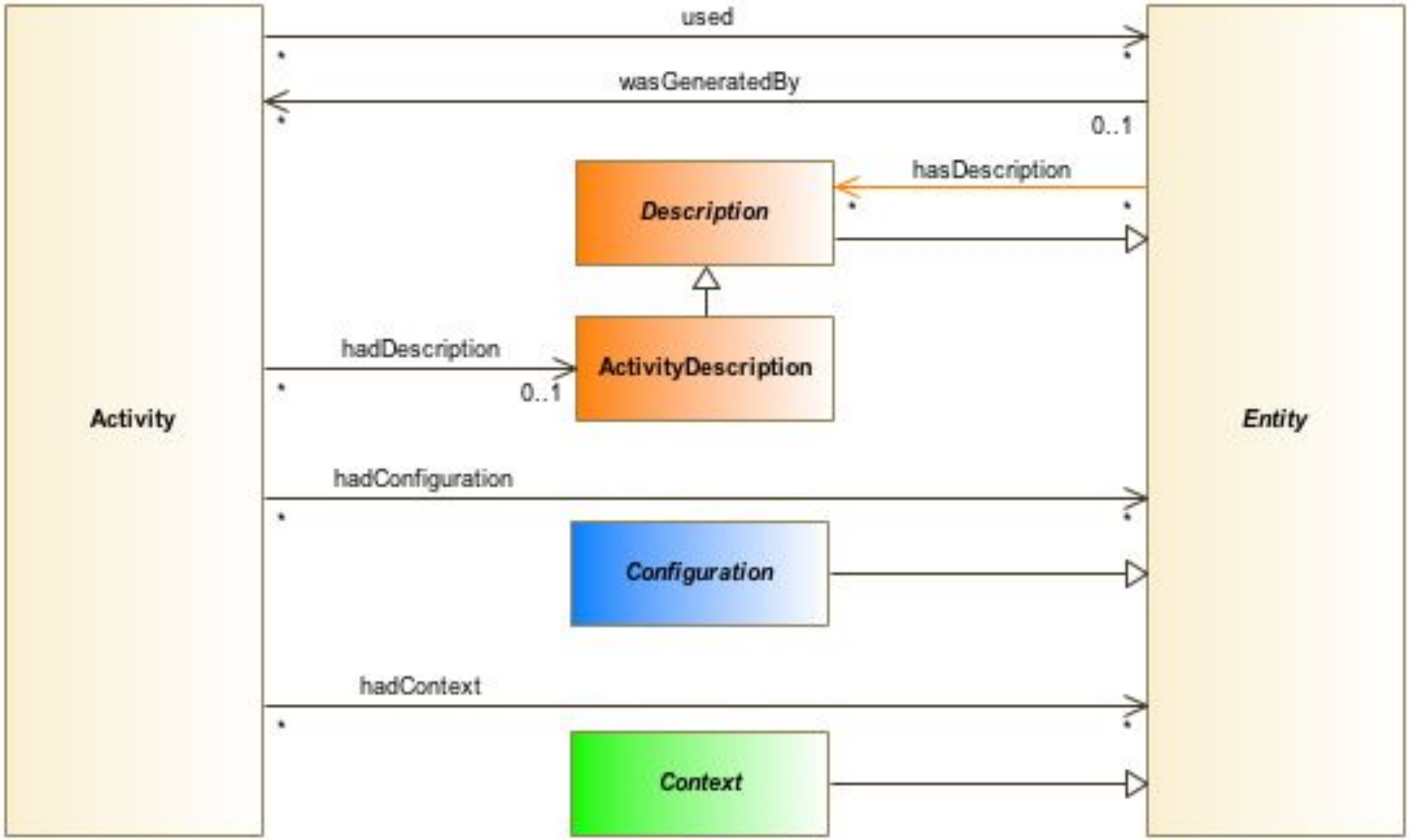
What was the weather during the observation, which hardware was used?

Context: information on the context that influences the development of an activity, but for which there is no or little control at the moment of its execution (e.g. Ambient Conditions, Instrumental Context, Execution Environment).

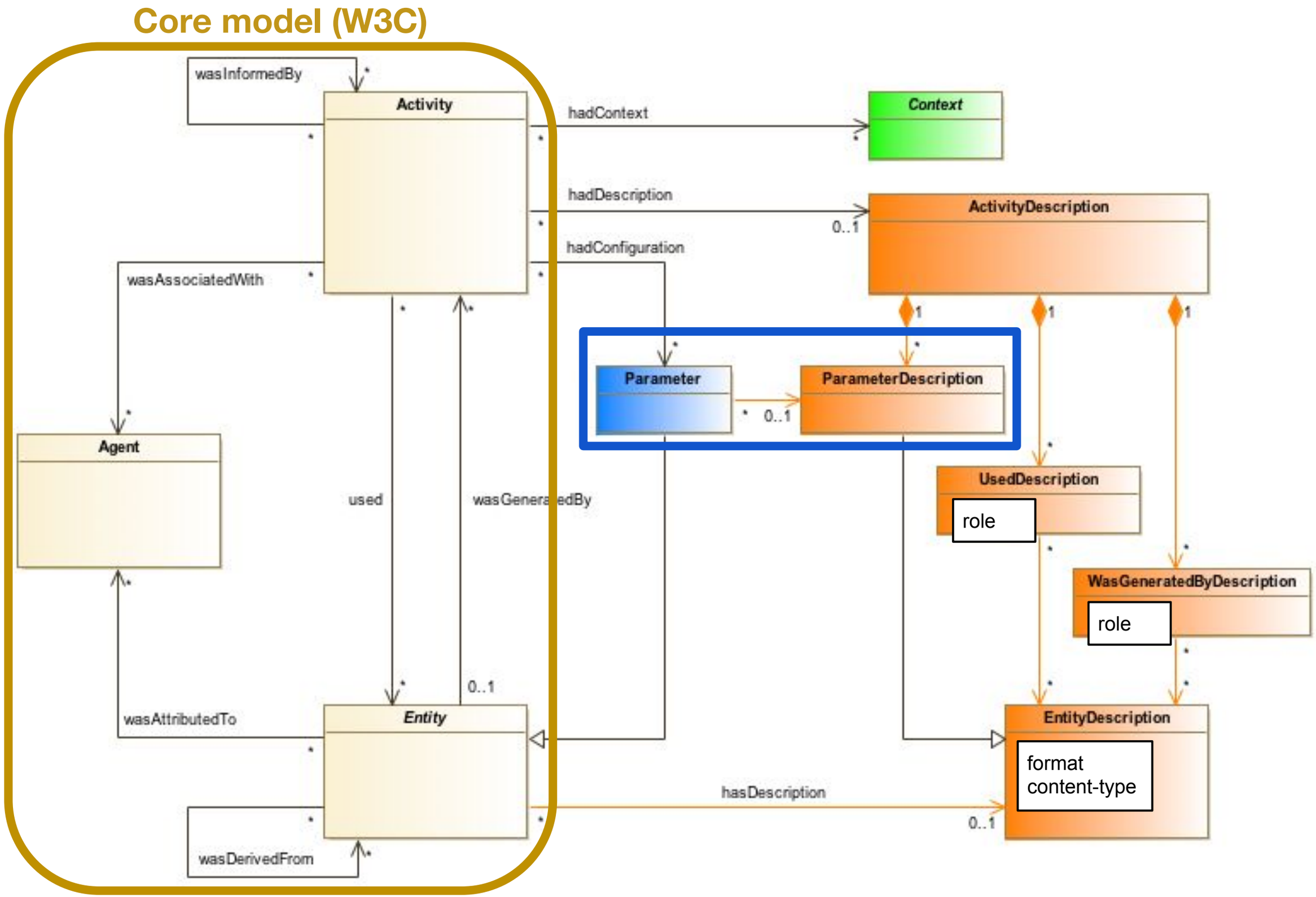
Specialized entities



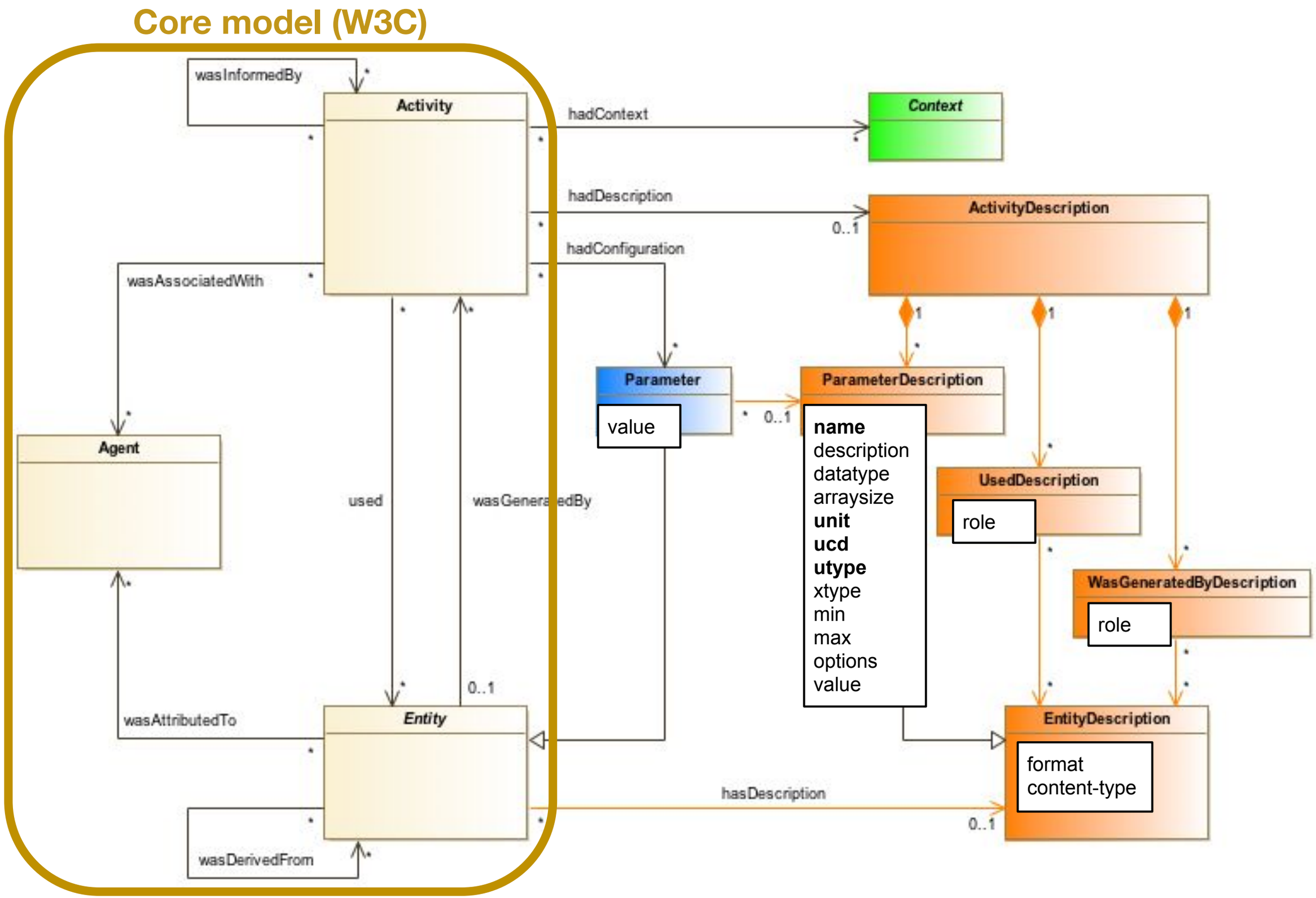
Specialized relations



Extended model general diagram



Extended model general diagram



ActivityDescrition examples

ESO recipe fors_img_screen_flat

ftp://ftp.eso.org/pub/dfs/pipelines/fors/fors-pipeline-manual-5.7.pdf

9.3.1 Input files

SCREEN_FLAT_IMG: required set of raw, unprocessed screen flat field frames.
MASTER_BIAS: required master bias frame. Just one should be given.

9.3.2 Output files

MASTER_SCREEN_FLAT_IMG: Master screen flat field calibration frame. Configuration parameters directly affecting this product are: --stack_method, --xradius, --yradius, -- degree, and --sampling.

9.3.3 Configuration parameters

The following parameters determine how the fors_img_screen_flat recipe will process the input frames.
--stack_method: Frames combination method. Default: average
See explanation in recipe fors_bias configuration parameters (Section 9.1.3, page 76).
--xradius: Median filter x radius (unbinned pixels). Default: 50 pixel
See the --yradius parameter.
--yradius: Median filter y radius (unbinned pixels). Default: 50 pixel
These parameters define the size of the running box used for smoothing the flat field for determining the large scale trend to remove. These parameters are ignored if the --degree parameter is greater than zero.
--degree: Degree of bivariate fitting polynomial. Default: -1
If this parameter is greater than or equal to 0, then a polynomial with the specified degree will be fitted to the illuminated part of the CCD for determining the flat field large scale trend to remove.
...

ActivityDescrition examples

Aladin hipsgen

Usage:

java -jar Aladin.jar -hipsgen in=file|dir [otherParams ... ACTIONs ...]

java -jar Aladin.jar -hipsgen -param=configfile

The config file must contain these following options, or use them directly on the command line :

Required parameter:

in=dir: Source image directory (FITS or JPEG|PNG +hhh or HiPS), unique image or HEALPix map file

Basic optional parameters:

out=dir: HiPS target directory (default ./+"AUTHORITY_internalID")

obs_title=name: Name of the survey (by default, input directory name)

creator_did=id: HiPS identifier (syntax: [ivo://]AUTHORITY/internalID)

hips_creator=name: Name of the person|institute who builds the HiPS

hips_status=xx: HiPS status (private|public clonable|clonableOnce|unclonable - default: public clonableOnce)

hdu=n1,n2-n3,...|all: List of HDU numbers (0 is the primary HDU - default is 0)

blank=nn: Specifical BLANK value

skyval=key|auto|%info|%min %max: Fits key to use for removing a sky background, or auto detection or percents of pixel histogram kept (central ex 99, or min max ex 0.3 99.7)

...

Advanced optional parameters:

hips_order=nn: Specifical HEALPix order - by default, adapted to the original resolution

hips_pixel_bitpix=nn: Specifical target bitpix (-64|-32|8|16|32|64)

...

ActivityDescrition examples

GammaPy image bin

Usage: gammapy image bin [OPTIONS] **EVENT_FILE REFERENCE_FILE OUT_FILE**
Bin events into an **image**.
You have to give the event, reference and out FITS filename.
Options:

--overwrite: Overwrite existing files?
-h, **--help**: Show this message and exit.

Submit your ADASS abstract

Web form inputs:

Title: Mr./Mrs./Ms./Mx./Dr./Prof.
First Name:
Last Name:
Name Printed on Name Badge:
University/Affiliation:
...
Banquet Dinner: yes/no
...
Abstract:

Output: registration email, records in a database

- Those activities define **configuration parameters** or **options**
- Some of those parameters are in fact (or point to) **main entities** that needs to be traced, e.g. **files**, **devices** or **documents**
- What is to be traced is the decision of the project or user that provides this activity
  → what is **relevant**?

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>
```

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 and next steps

- W3C core model **robust** and **generic** (+external compatibility)
- IVOA model adds **relevant** provenance information
- Working **implementations** from different projects and institutes
- RFC comments:
  - keep only the generic model?
  - reduce length
  - more normative
  - less features? (cost for implementers)
  - inconsistencies?
- Keep in mind that all this **is** working...