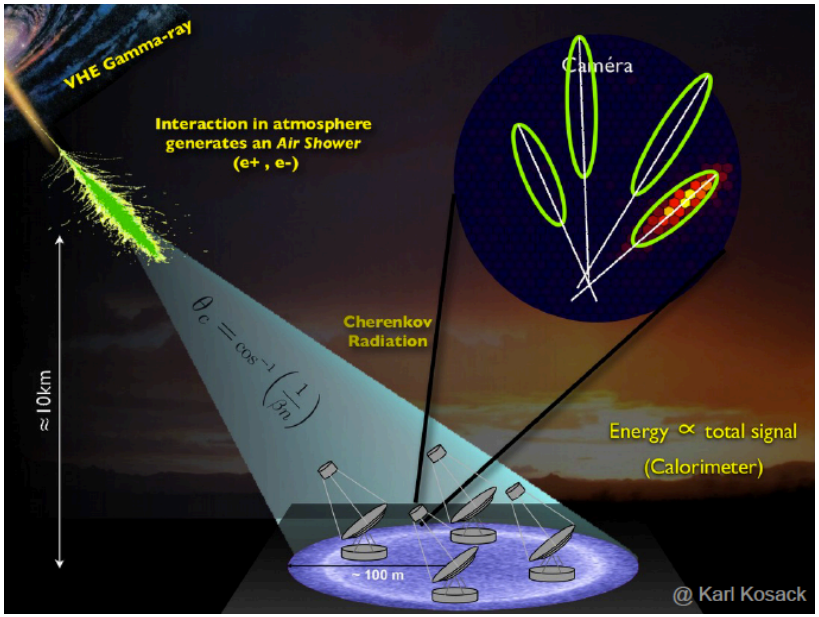
IVOA PROVENANCE DATA MODEL CTA & POLLUX use cases

Michèle Sanguillon[(*)(1)(2)], **Mathieu Servillat**[(**)(1)], **Julien Le Faucheur** [(**)(1)], **Catherine Boisson**[(**)(1)], **Johan Bregeon**[(*)(1)], **Ana Palacios**[(*)(2)], **Mireille Louys**[(***)(3)], **François Bonnarel**[(***)(3)], **Pierre Le Sidaner**[(****)(3)]

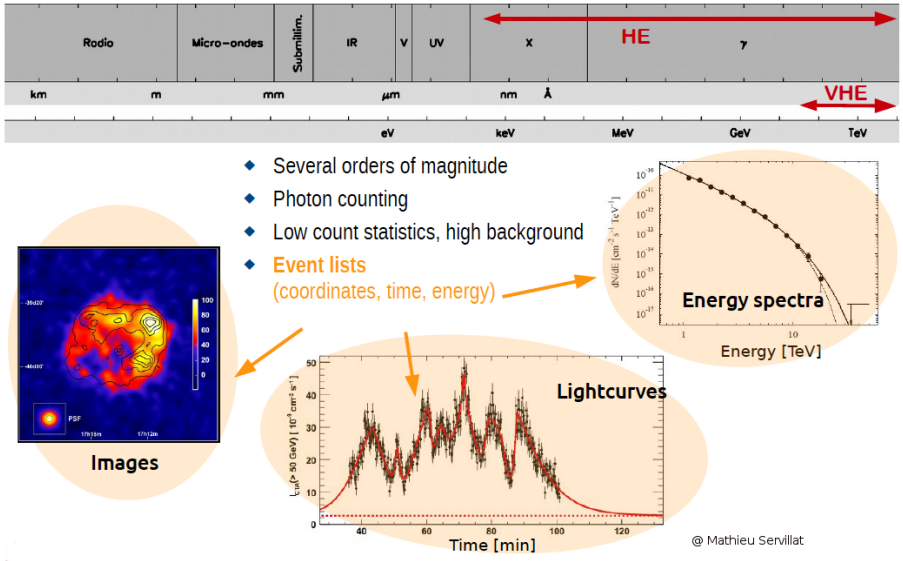
(*) LUPM, Montpellier (**) LUTH, Meudon (***) CDS, Strasbourg, (****) DIO, Paris
(1) CTA, (2) Pollux, (3) VO Experts

CTA Context



- Very high energy gamma ray instrument
- 3 types of telescopes in CTA
- Complex data :
  - Indirect detection
  - Need simulations to compare acquired data to expected ones
- Final products data available on the VO

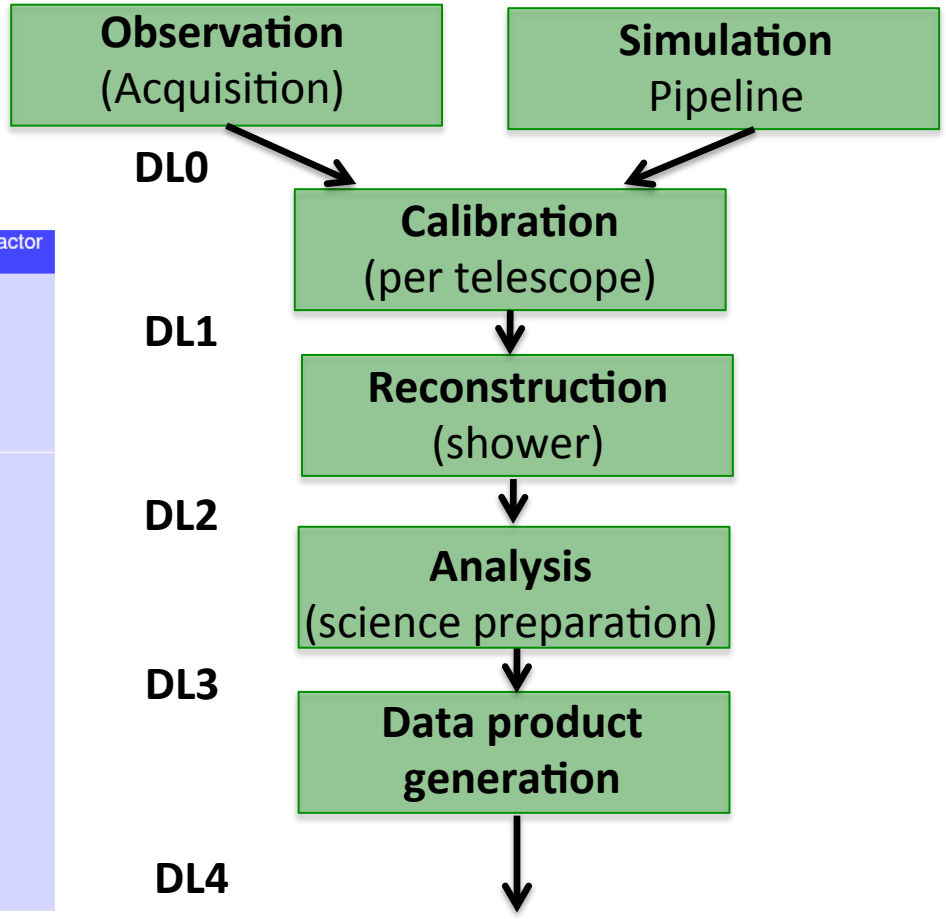
CTA will be the first Cherenkov Observatory providing its high level data (event lists, spectra, sky maps) on the Virtual Observatory



CTA data and workflows

Different levels of data : DL0 to DL5.
DL3, DL4 and DL5 data available on the VO

Data Level	Short Name	Description	Data reduction factor
Level 0 (DL0)	DAQ-RAW	Data from the Data Acquisition hardware/software.	
Level 1 (DL1)	CALIBRATED	Physical quantities measured in each separate camera: photons, arrival times, etc., and per-telescope parameters derived from those quantities.	1-0.2
Level 2 (DL2)	RECONSTRUCTED	Reconstructed shower parameters (per event, no longer per-telescope) such as energy, direction, particle ID, and related signal discrimination parameters.	10^{-1}
Level 3 (DL3)	REDUCED	Sets of selected (e.g. gamma-ray-candidate) events, along with associated instrumental response characterizations and any technical data needed for science analysis.	10^{-2}
Level 4 (DL4)	SCIENCE	High Level binned data products like spectra, sky maps, or light curves.	10^{-3}
Level 5 (DL5)	OBSERVATORY	Legacy observatory data, such as CTA survey sky maps or the CTA source catalog.	10^{-5} - 10^{-3}



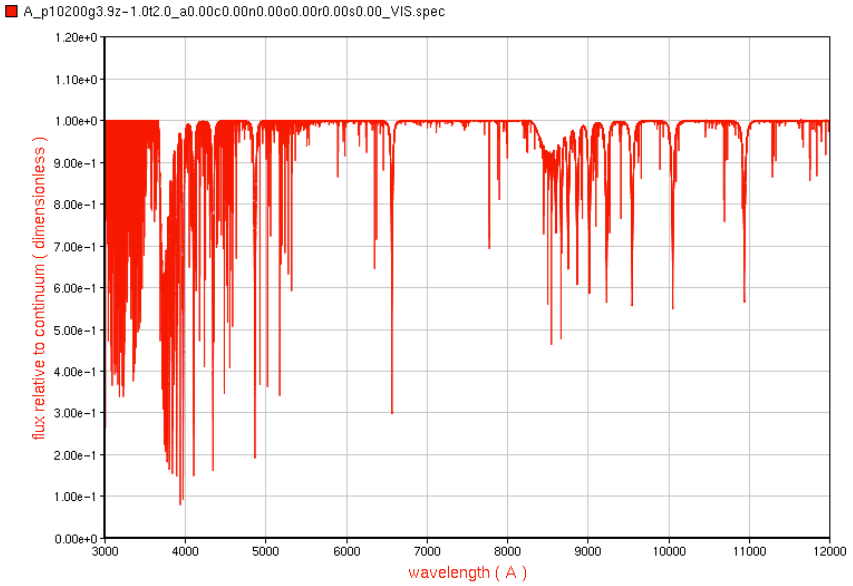
CTA use cases

- **Use case 1:**
  Search data available for a given Target at a given time interval. → ObsCore, Any protocol
- **Use case 2:**
  Search public data for all blazars → Extended ObsCore, TAP
- **Use case 3:**
  Search data that include LST (Large Size Telescope). → CTA ObsConfig, TAP
- **Use case 4:**
  search data produced using a given reconstruction method → No data model, No current protocol
- **Use case 5:**
  search data for a given target produced with loose cuts → No data model, No current protocol

Pollux Context

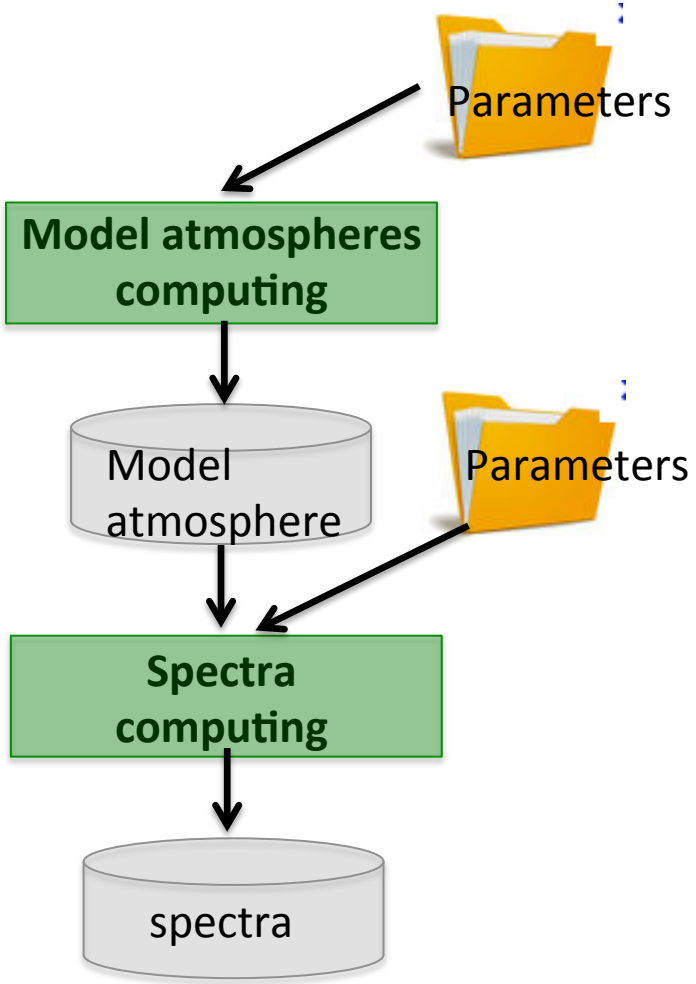
temperature = 5000°K
gravity = …
…

POLLUX database collects and presents synthetic spectra computed* at high resolution.



* The computing of the spectra is done by the producers and not done on the fly.

Pollux data and workflow



Only the spectra are available on the VO

Pollux use cases

- **Use case 1:**
  Show me a list of synthetic spectra satisfying :
  - domain of wavelength = visible
  - domain of effective temperature = [4000, 5000]

- **Use case 2:**
  Show me a list of synthetic spectra satisfying :
  - code for model atmosphere = MARCS
  - type of model atmosphere = spherical

- **Use case 3:**
  Show me a list of synthetic spectra satisfying :
  - code for spectral synthesis = turbospectrum
  - version of this code = 2008.1

Data Model:
- Obs* could not be applied
- SimDM implements only a simulator and a PostProcessor

Protocol
- SSA protocol with format = METADATA but only few criteria are currently available.
- SimDAL ?

Other use cases

- Concerning the observed spectra, the provenance of them is important and the provenance characteristics are mostly described by the ObsConfig part of the spectral data model. But all observed spectra offered to users are not raw data. They have often been transformed by programs (calibration, ...). No processing provenance information is given about those programs.

- **Use case :**

  For a given star identified by POS and SIZE, show me a list of spectra satisfying :
  - Stokes parameter = Q
  - Result of the LSD (code 1) processing = definite

Provenance in the IVOA

- **Explains how data sets were produced:**
  - Observing process and conditions
  - Data reduction, selection and extraction methods applied to raw measures to build up science-ready data products (source lists, spectra, light curves, images, …)
  - Worflows to build theorical data (spectra, images, …)
- **Helps VO users to:**
  - Derive selection criteria to filter out suitable data for his/her scientific needs
  - Estimate better which data release fits the best for their needs
  - Run his/her own reduction method on intermediate data products in order to refine data analysis

TO DO

- **VO users need :**
  - **to know what we are talking about**
  - **to know how data sets were produced**
  - **To select data on provenance criteria**

- **We have to define/identify:**
  - **a Data Model**
  - **a way to publish the provenance information**
  - **a format of description to describe all the provenance information**
  - **a protocol to query**

Provenance in the W3C

W3C Provenance definition:

« Provenance is information about entities, activities and people involved in producing a piece of data or thing, which can be used to form assesments about its quality, reliability and trustworthiness. PROV-DM is the concepual data model that forms a basis for the W3C provenance (PROV) familiy of specfications. »

4 recommendations (30/04/2013)

PROV-DM: the PROV data model

PROV-O: the PROV ontology

PROV-Constraint: Constraints of the PROV Data Model

PROV-N: a notation for provenance aimed at human consumption

and a number of non-prescriptive notes

PROV-XML: an XML schema for the PROV data model

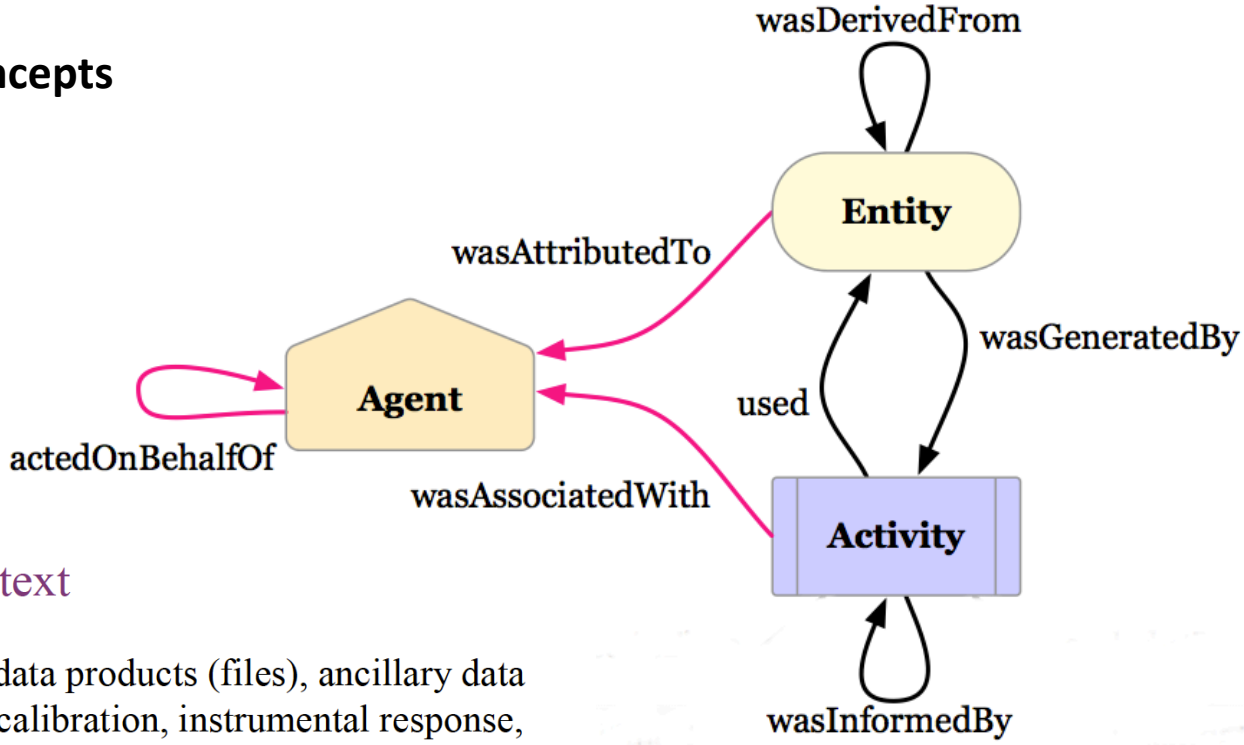
PROV-AQ: Provenance access and query

Benefits:

- 4 recommendations and a number of non-prescriptive notes
- Tools to validate and translate a description format in another
- Possible to define our own attributes

W3C Provenance Data Model

Core Concepts



In our context

Entity
- data products (files), ancillary data (calibration, instrumental response, etc.), processing parameter files

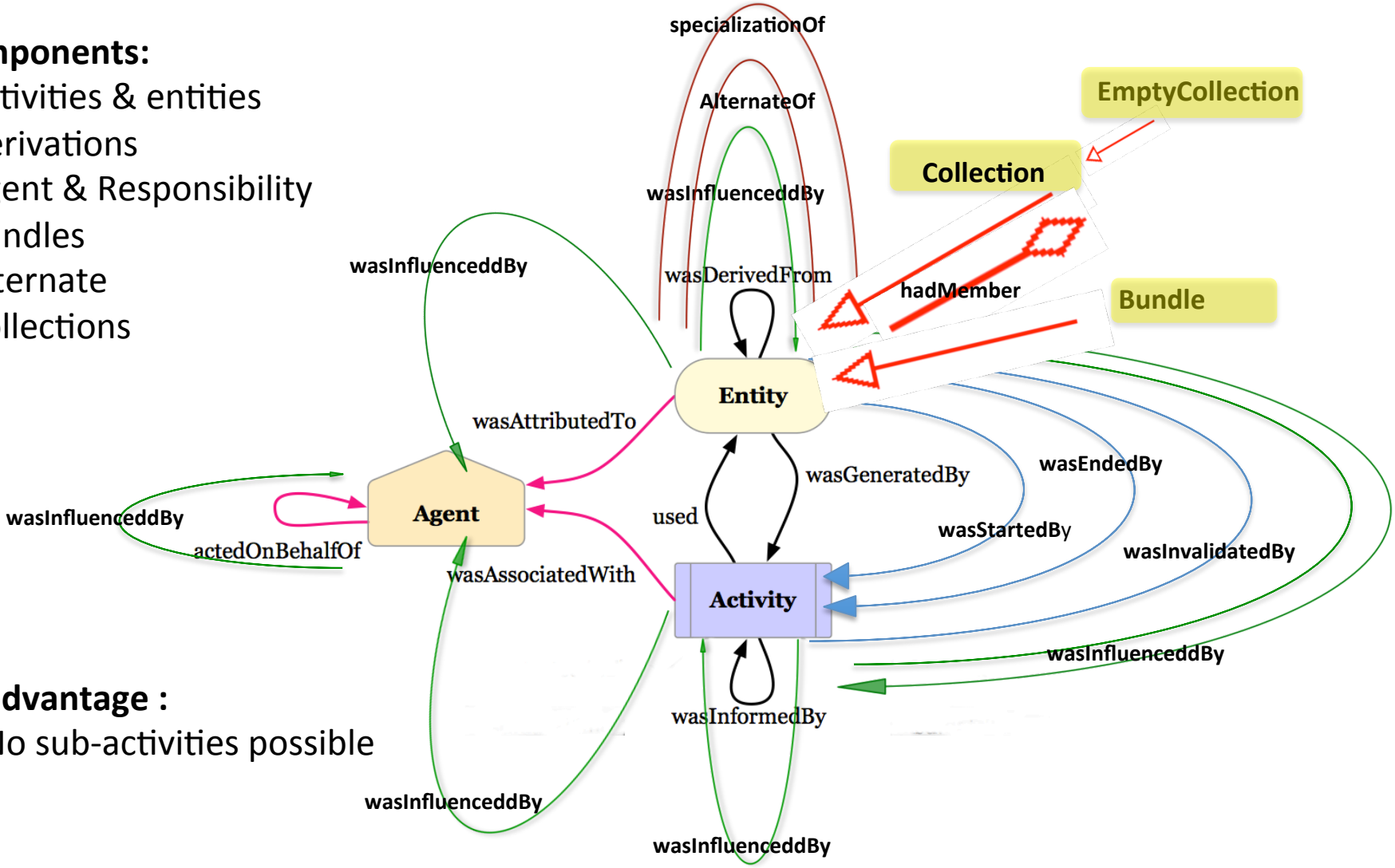
Activity
- data acquisition, mosaicing, regridding, fusion, calibration, …, transformation

Agent
- Telescope astronomer, pipeline operator, principal investigator, etc.

W3C Provenance Data Model

6 components:

- Activities & entities
- Derivations
- Agent & Responsibility
- Bundles
- Alternate
- Collections

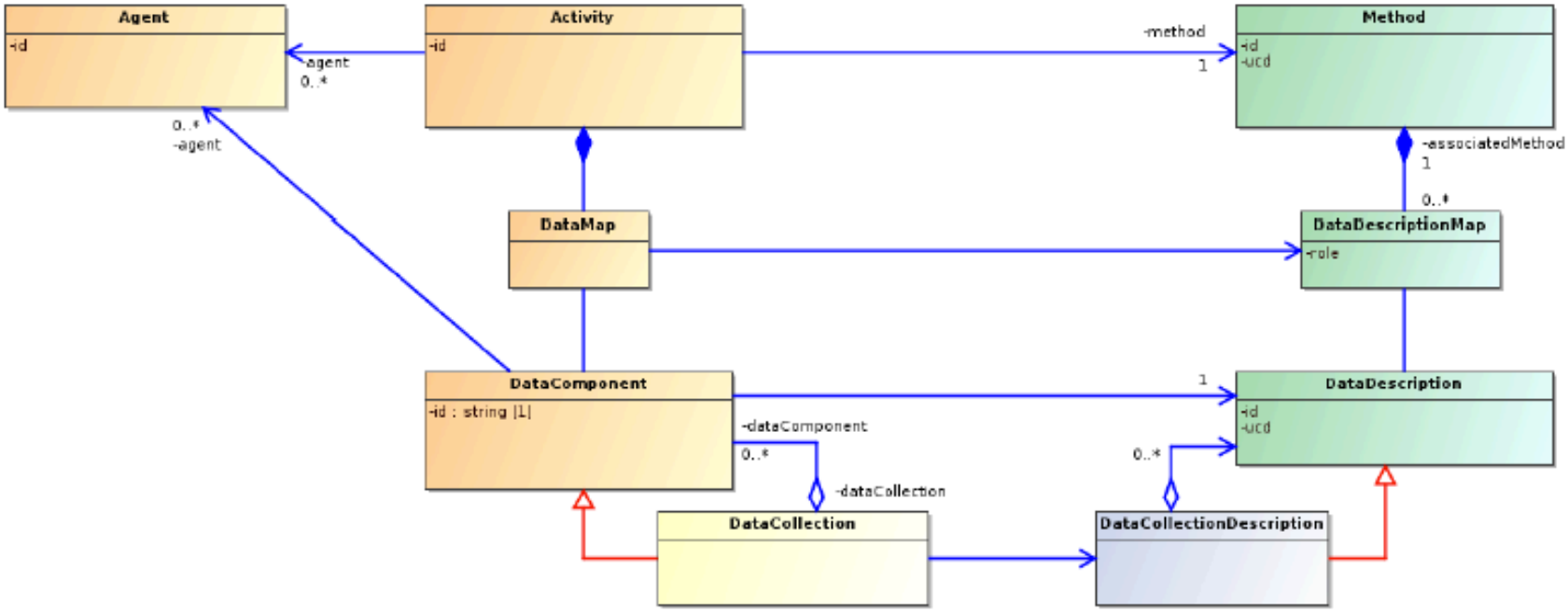


Disadvantage :

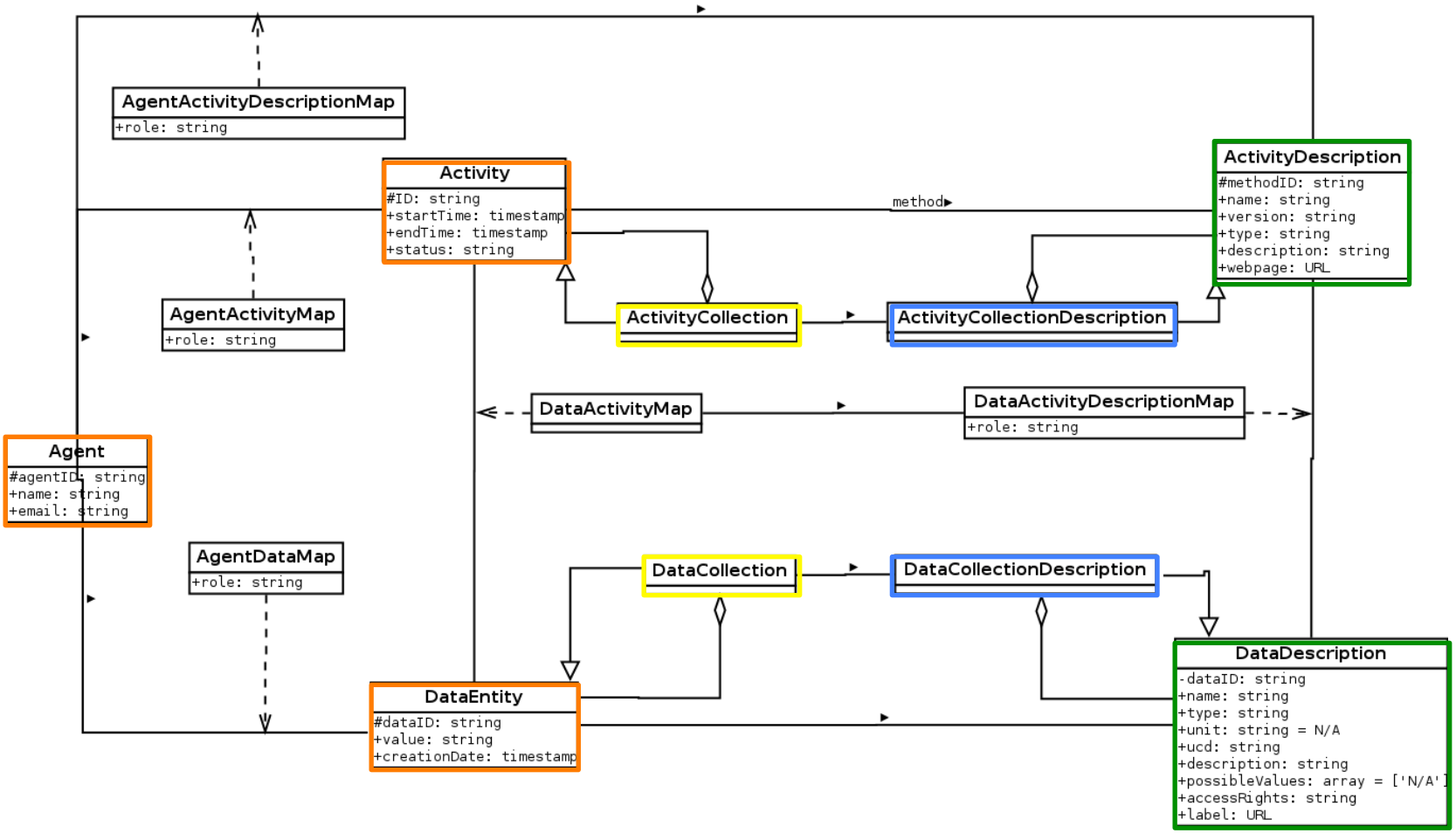
- No sub-activities possible

IVOA current DM

IVOA Provenance Data Model
Version 0.1
IVOA Working Draft 2015-05-18



IVOA proposal DM



Provenance information creation

- VO published data files
  - We want to describe their provenance
  - This provenance set could be created at the same time of the file
- What information ?
  - Activities & entities & relations
  - Agents ?
- What level of detail ?
  - A set of input parameters contain important parameters for the VO user and not important ones
  - If we want to describe important parameters, the parameter files could be described as a collection of data and each iportant parameter as a data

A way to publish and a format of description

- Dedicated file (DATALINK)?
  - Dedicated service
  - Format = PROV-N, JSON, PNG, XML, VOTable (TBD)
- File header?
  - FITS : Keywords (history) ?
  - Txt : PROV-N, JSON, … ?
  - VOTable (TBD)

Query

- SSA
  - with FORMAT=METADATA
  - with PROV criteria
- TAP
  - Extension with ProvTAP ?
- SIA, SLA
  - Extensions ?

A beginning of implementation

POLLUX

- Provenance information creation
- Script which generates for each spectrum a provenance file in JSON format

```
Teff        = '14800'              / effective temperature (K)     - model atmosphere data
logg        = '4.1'                / log10(gravity) (cgs)         - model atmosphere data
```

```
"entity": {
....
"pollux:14800.mod_2012_Teff" : {"voprov:type": int, "prov:value":"14800",
            "voprov:unit": "K, "voprov:ucd":"phys.temperature.effective",
            "voprov:description":"effective temperature (K)     - model atmosphere data"} ,
"pollux:14800.mod_2012_logg" : {"voprov:type": float, "prov:value":"4.1",
            "voprov:unit": "log(cm/s2), "voprov:ucd":"phys.gravity;arith.zp",
            "voprov:description":"log10(gravity) (cgs)         - model atmosphere data"} ,
 ... }
```

In the next future

POLLUX

- Provenance information creation

  Script which generates for each spectrum a provenance file in other format (PROV-N, PNG, XML, VOTable)

- Provenance information publication
  - Service which allows to download the provenance file
  - DATALINK added for each spectrum in the SSA response
- Provenance query

THANK YOU
FOR YOUR ATTENTION