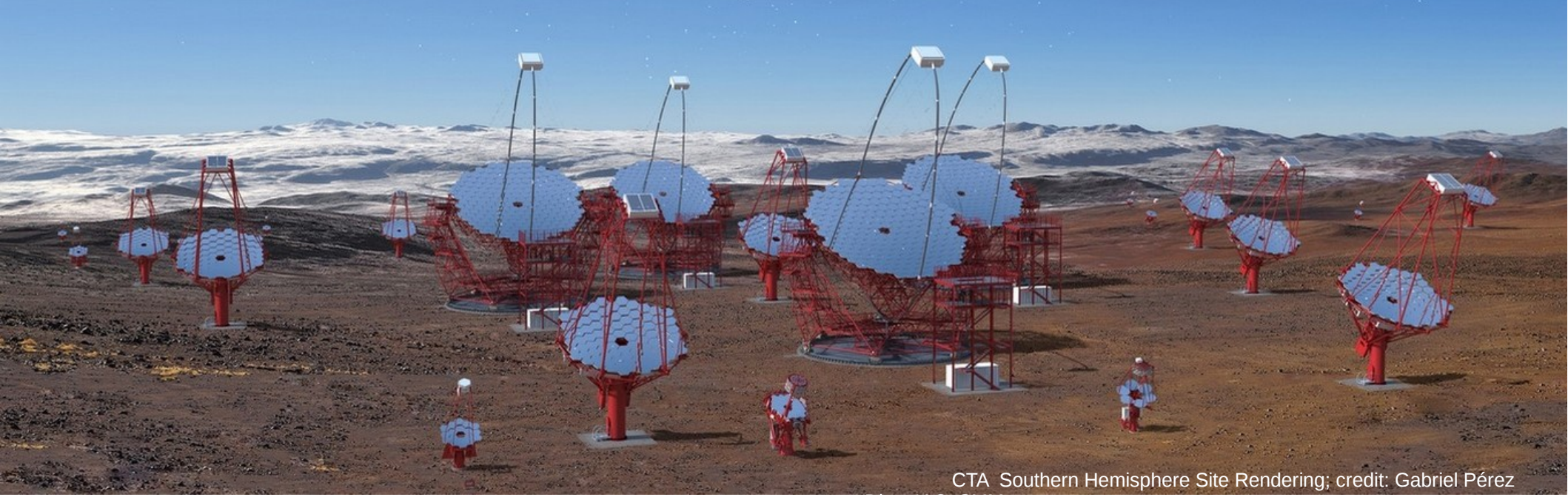
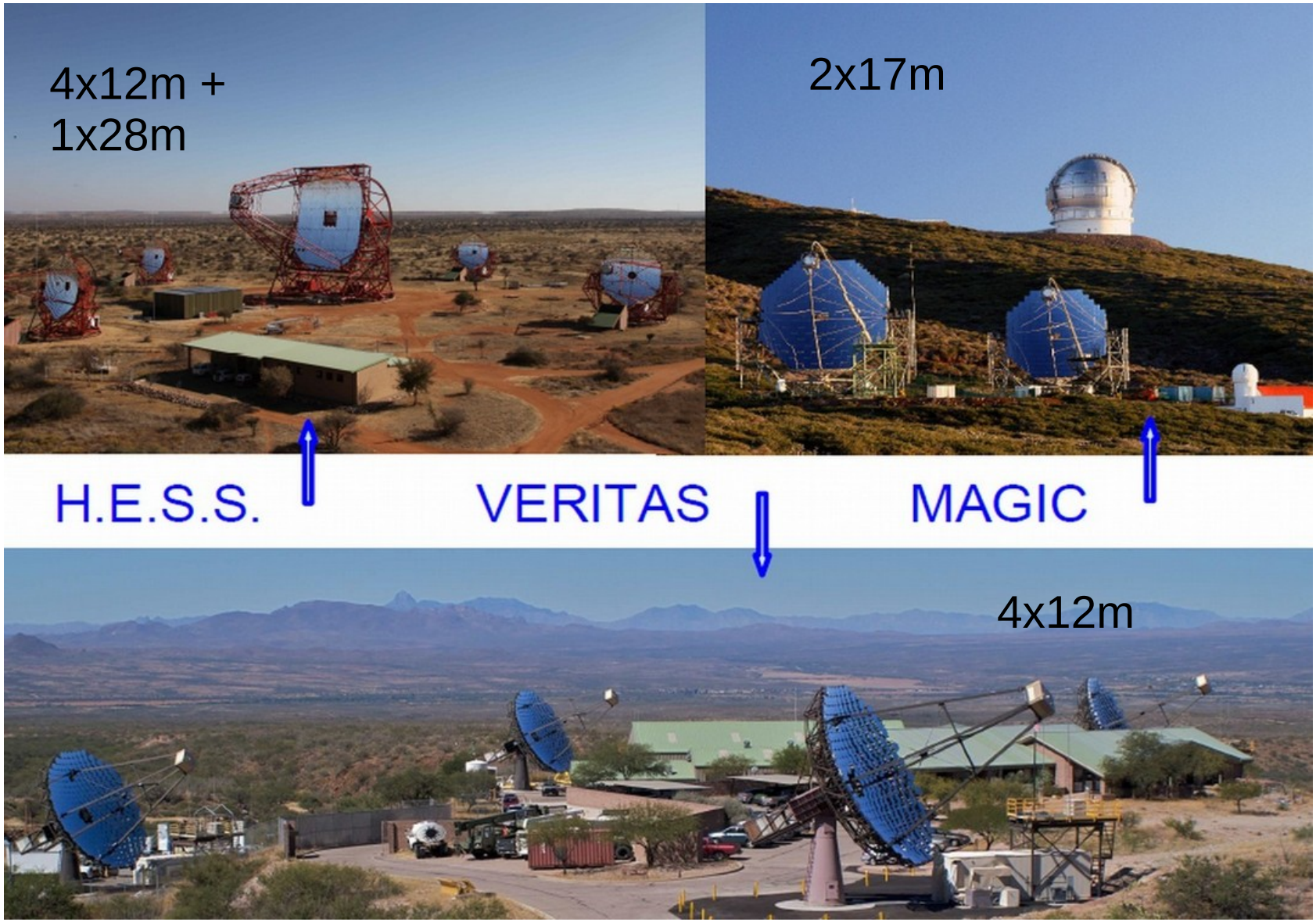
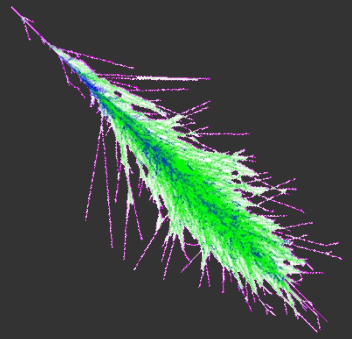

CTA Southern Hemisphere Site Rendering; credit: Gabriel Pérez

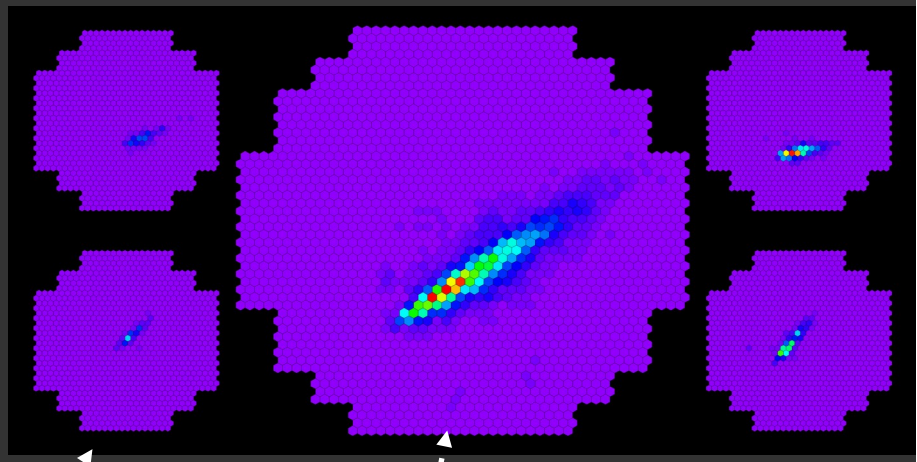
Provenance and data access in the context of Cherenkov astronomy

C. Boisson & M. Servillat
LUTh, Observatoire de Paris
European Data Provider Forum, Heidelberg June 2018

Ground based IACTs







Dark nights → small duty cycle

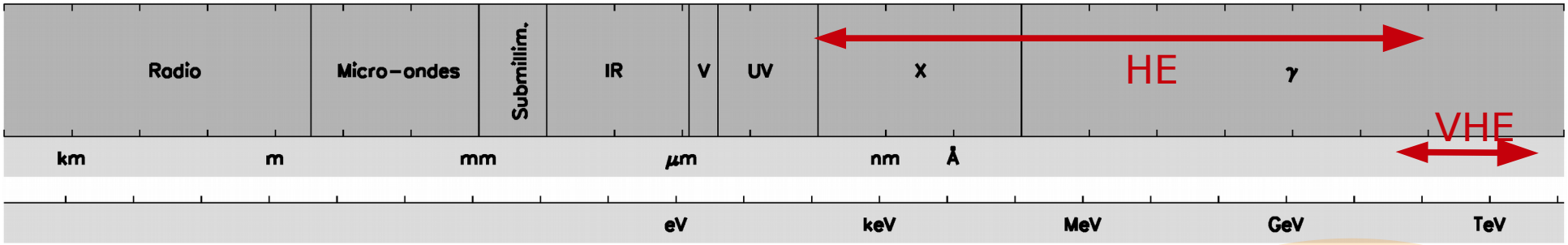
Event reconstruction :

photon, particle shower, Cherenkov light (faint, few nanoseconds)

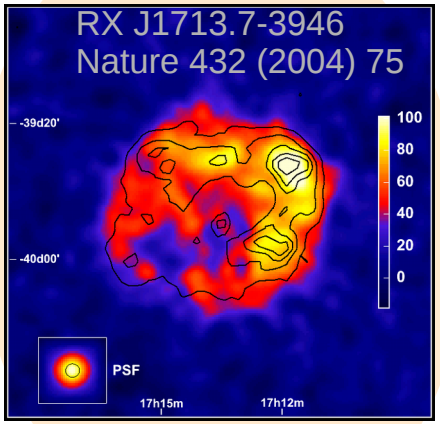
Atmosphere = calorimeter
Simulations, assumptions

Complex metadata :
need to be structured

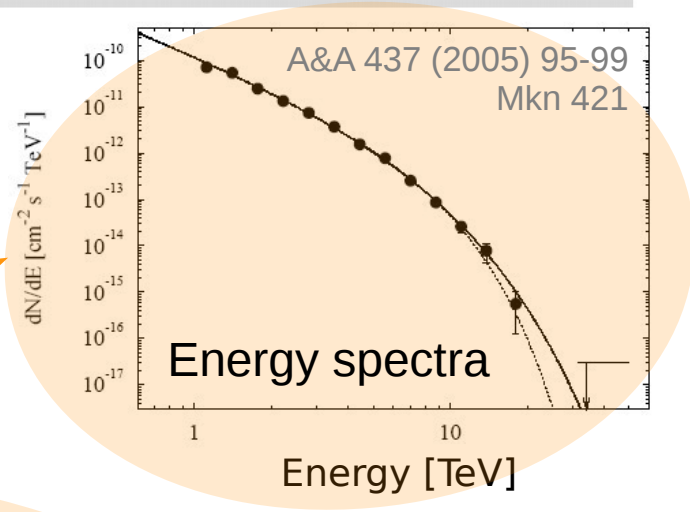
Very high energy data

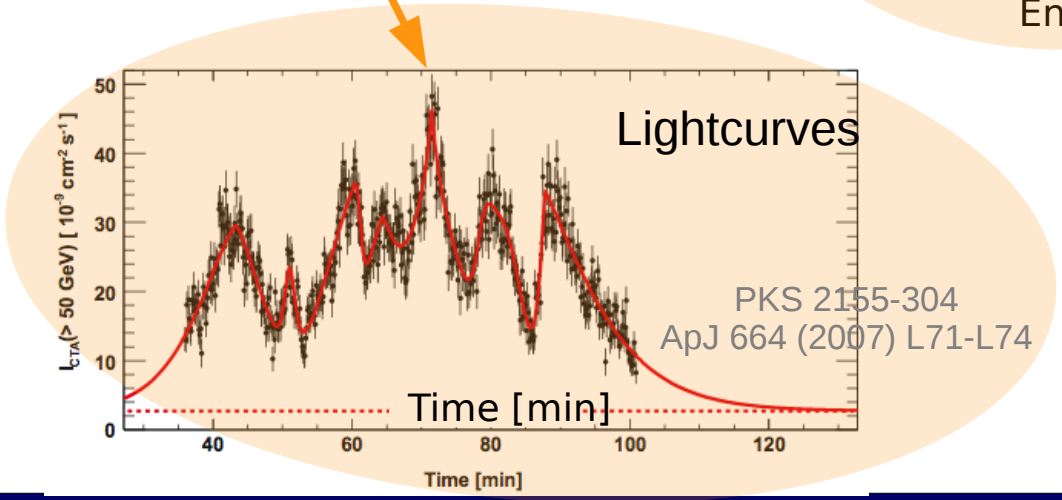


- Several orders of magnitude
- Photon counting
- Low count statistics, high background
- **Event lists** (coordinates, time, energy)



Images





@ M. Servillat

H.E.S.S. AGN

Only a few hours of useful data summed over a long time

Not pixels but assymetric energy bins

Close

HESS J0152+017

Observation									Curation								
name	comments	pointing alpha	pointing alpha sys	pointing alpha stat	pointing delta	pointing delta sys	pointing delta stat	publisher	curation date	version	rights	contact name	contact email	title	creator	creation date	creation type
HESS J0152+017	October to November 2007 summed data; significance	1:52:33.500	1.3	5.3	1:46:40.200	20	107	VO-Paris	02-08-2008	1.0	Public	C. Boisson	catherine.boisson@obspm.fr	Extragalactic	C. Boisson	28-07-2008	Archival

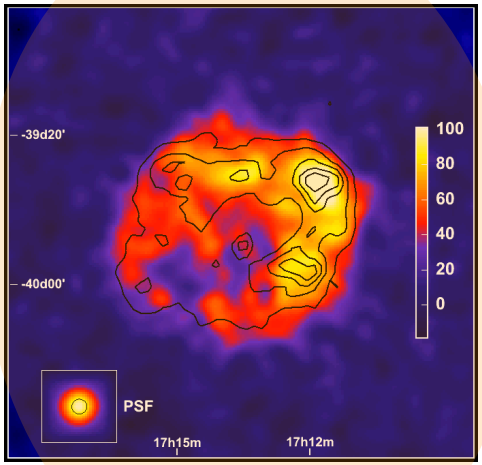
Time Axis			Spectral Axis	Spectral Data (E in TeV)	Flux Data (dN/dE in cm-2.s-1.TeV-1)	
bounds start	bounds stop	livetime	energy threshold	value	value	stat error
52412.075	53504.895	14.7	0.3	0.308477	[illegible]	6.67404e-12
				0.484509	5.64448e-12	1.63425e-12
				0.760992	1.20326e-12	5.01844e-13
				1.19525	3.12378e-13	2.16144e-13
				1.87731	1.18592e-13	8.50016e-14
				2.9486	4.0974e-14	3.51122e-14

Segment										Quality	Cuts		Background	
length	data type	imgfile	comments	background	hypothesis power law	hypothesis gamma	hypothesis ngamma	hypothesis chi2	hypothesis dof	mean zenith angle	name	description	name	description
6	Spectrum		Aharonian et al., A&A 481 (2008) L103	Reflected model	Single	2.95	173	2.16	4	26.9	Hillas soft cuts	Soft Cuts: as standard cuts but optimized for a 1% Crab flux (>100 GeV) source with a photon index of 5.0. * a 5/10 cleaning * a charge cut at 40 p.e. * a nominal distance cut at 2 degrees * a Mean Scaled Width between -2 and 0.9 * a Mean Scaled Length between -2 and 1.3 * a Theta^2 cut of 0.02	Reflected model	Technique used in standard wobble observation mode. See Aharonian et al. (H.E.S.S. Collaboration), A&A 457, 899 (2006)

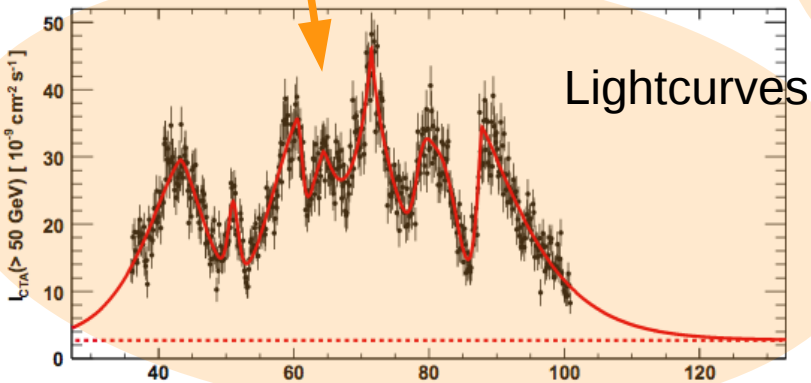
Close

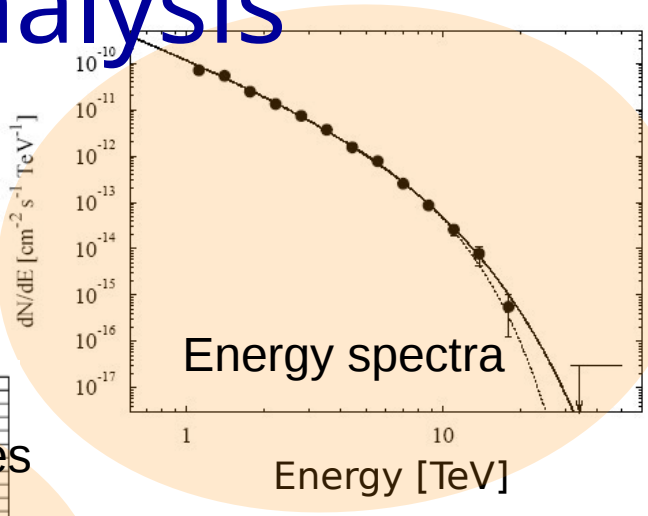
Multi-wavelength analysis

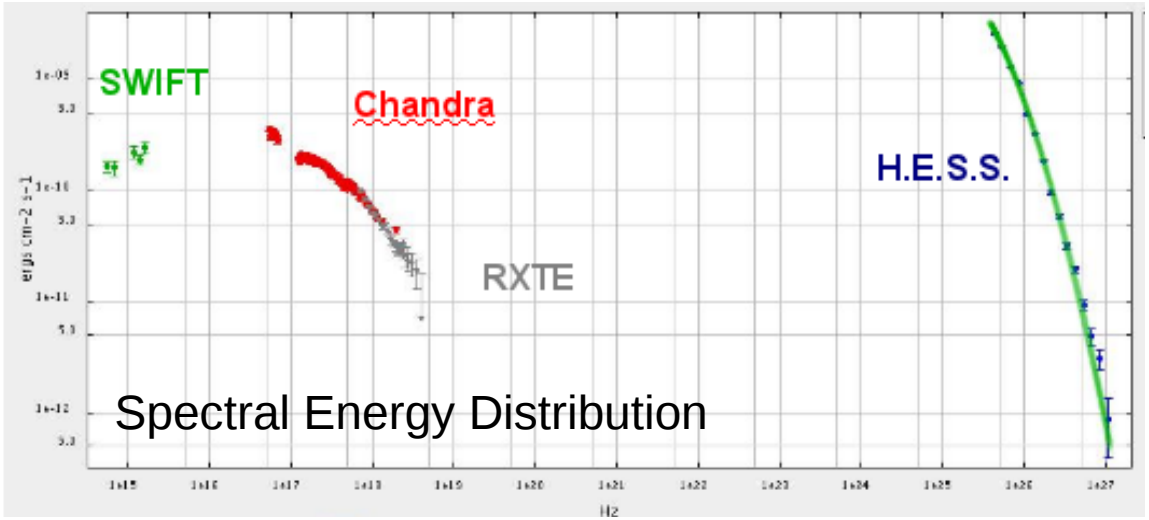
Event lists
(coordinates, time, energy)



Images



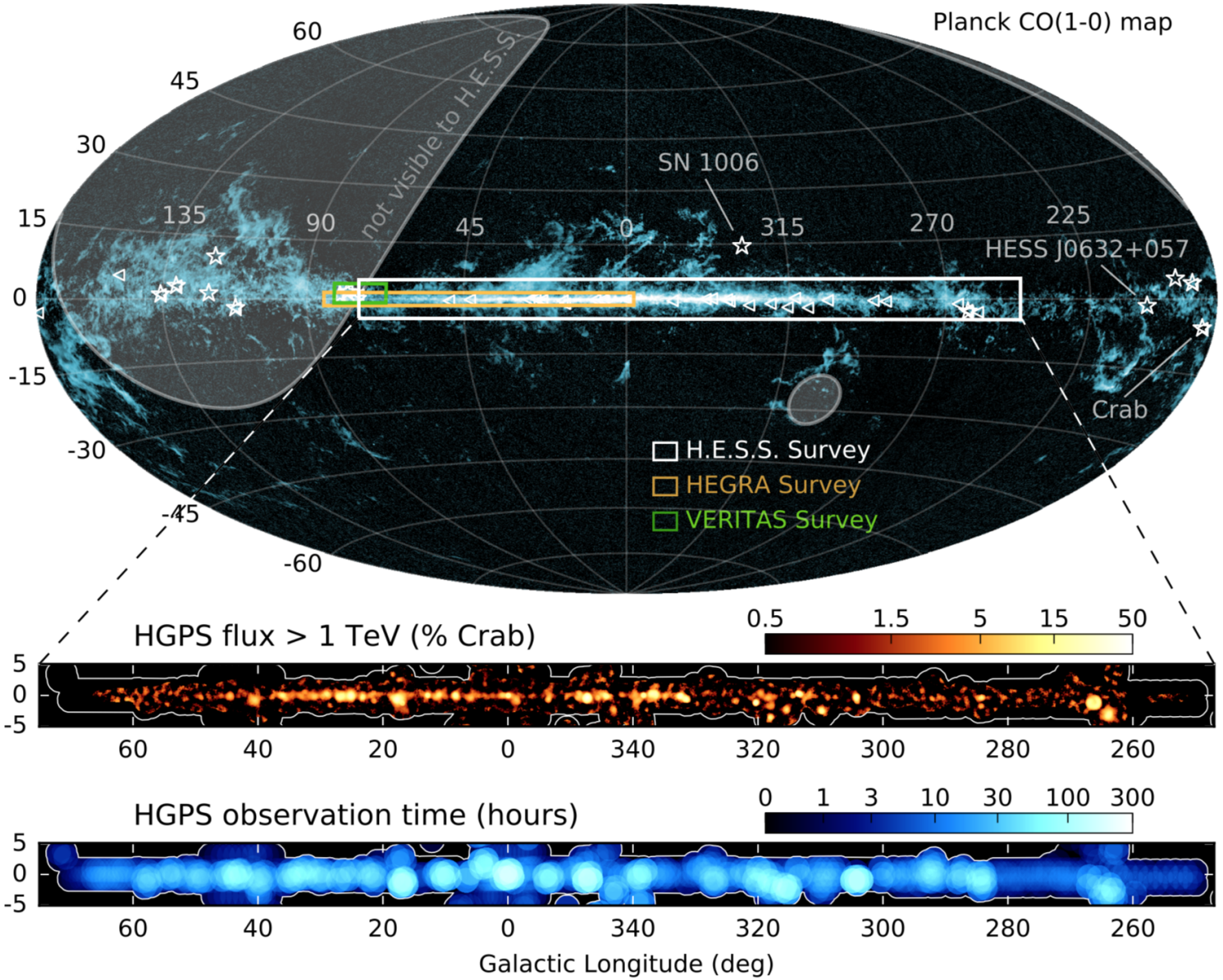




Compatible data at other wavelength?

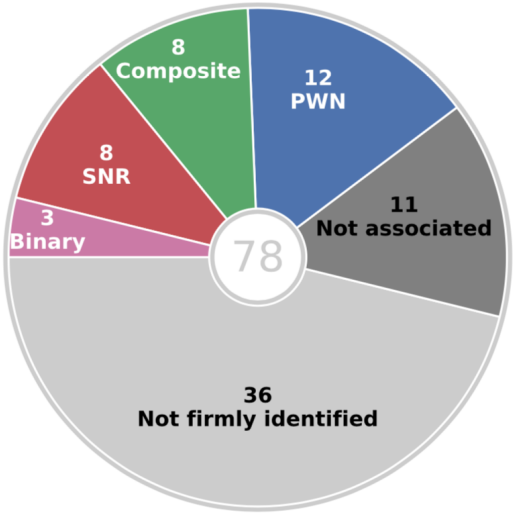
Simultaneous
Calibrated
Specific Processing?
Context?

H.E.S.S. Galactic plane survey



3000 hr of observations, sensitivity better than 2% of Crab nebula flux

extended and point-like sources



HESS Collab., A&A 612, A1 (2018)

Northern and Southern Hemisphere Site Rendering; credit: Gabriel Pérez Diaz, IAC, SMM

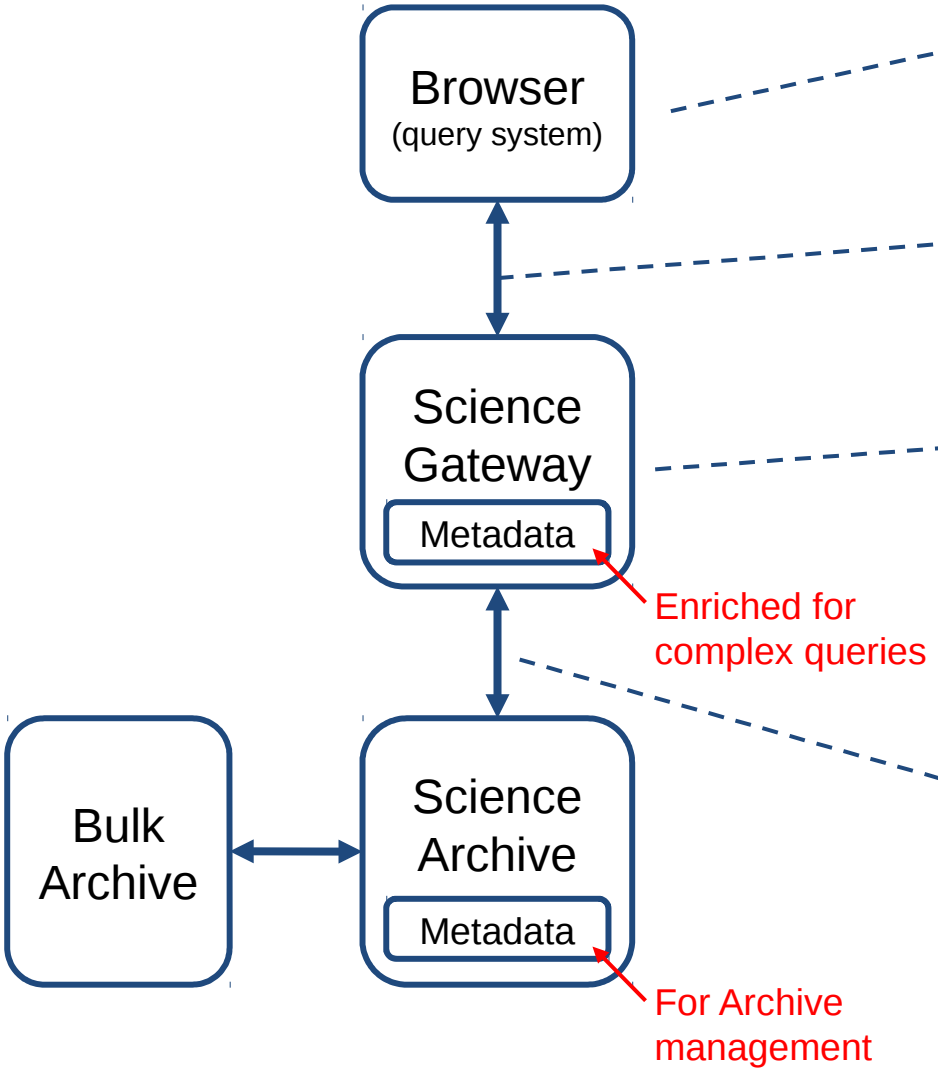
CTA data access use cases

- The **PI** of a successful proposal wants to retrieve the data
  - **Simple query** by obs_id (or PI name, or direct link sent to the PI)
  - Need user authentication and authorization
- A CTA Science User wants to find a **specific data set**
  - **Complex query**
  - Using Cone Search (RA, Dec) and/or other information (time range, spectral range, instrument configuration, nature of the target, keywords in the proposal, data processing details, …)
- A Science User wants to gather more information on a source **detected at other wavelengths**
  - No knowledge about CTA a priori
  - **Query limited to "generic" information** sent to several archives

⇒ The Virtual Observatory (VO) framework is useful for all those use cases

Science Gateway in the VO framework

CTA Data Access



In the Virtual Observatory Framework

Client: submits a query
- **VO tools** (Topcat, Aladin, scripts…)
- Dedicated **Web Client**

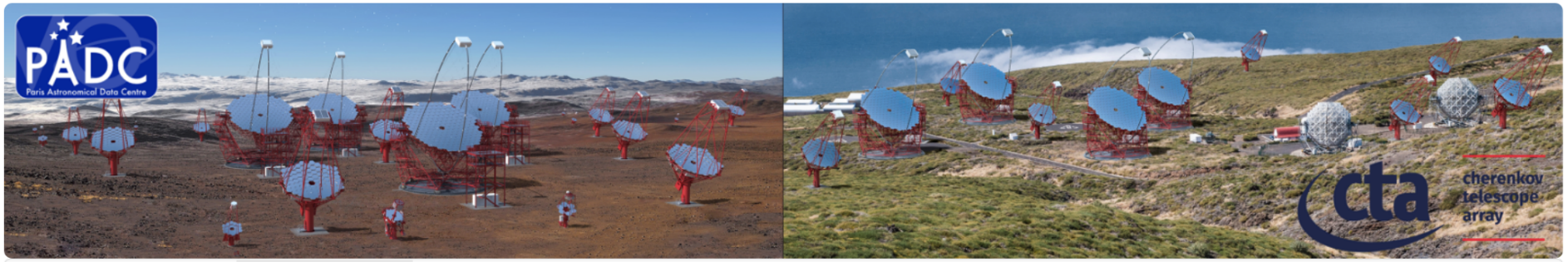
Protocol: standard for query exchange
- **ADQL** (Astronomical Data Query Language)
- **TAP** (Table Access Protocol)

Server: computes query results
- **TAP** Service
- VO **Data Models** (ObsCore, DataSet, ...)
  - RA → s_ra
  - Dec → s_dec
  - obs_id, t_min, t_max, access_url, …
- ⇒ **ObsTAP** Service

Retrieval System:
- VO ObsCore **access_url** + **DataLink**
- Any service at the **access_url**
  - FTP, HTTP server
  - VO Space
- e.g. https://archive.cta.org/retrieve?id=###

CTA Data Distiller

https://voparis-cta-test.obspm.fr



CTA Data Distiller | Search Form | Results | Sign in

Cone Search

Target Name	PKS 2155-304	Used to query Simbad with Sesame and set RA/Dec.
Source RA (deg)	329.717	Right Ascension.
Source Dec (deg)	-30.226	
Search radius (deg)	0.001	

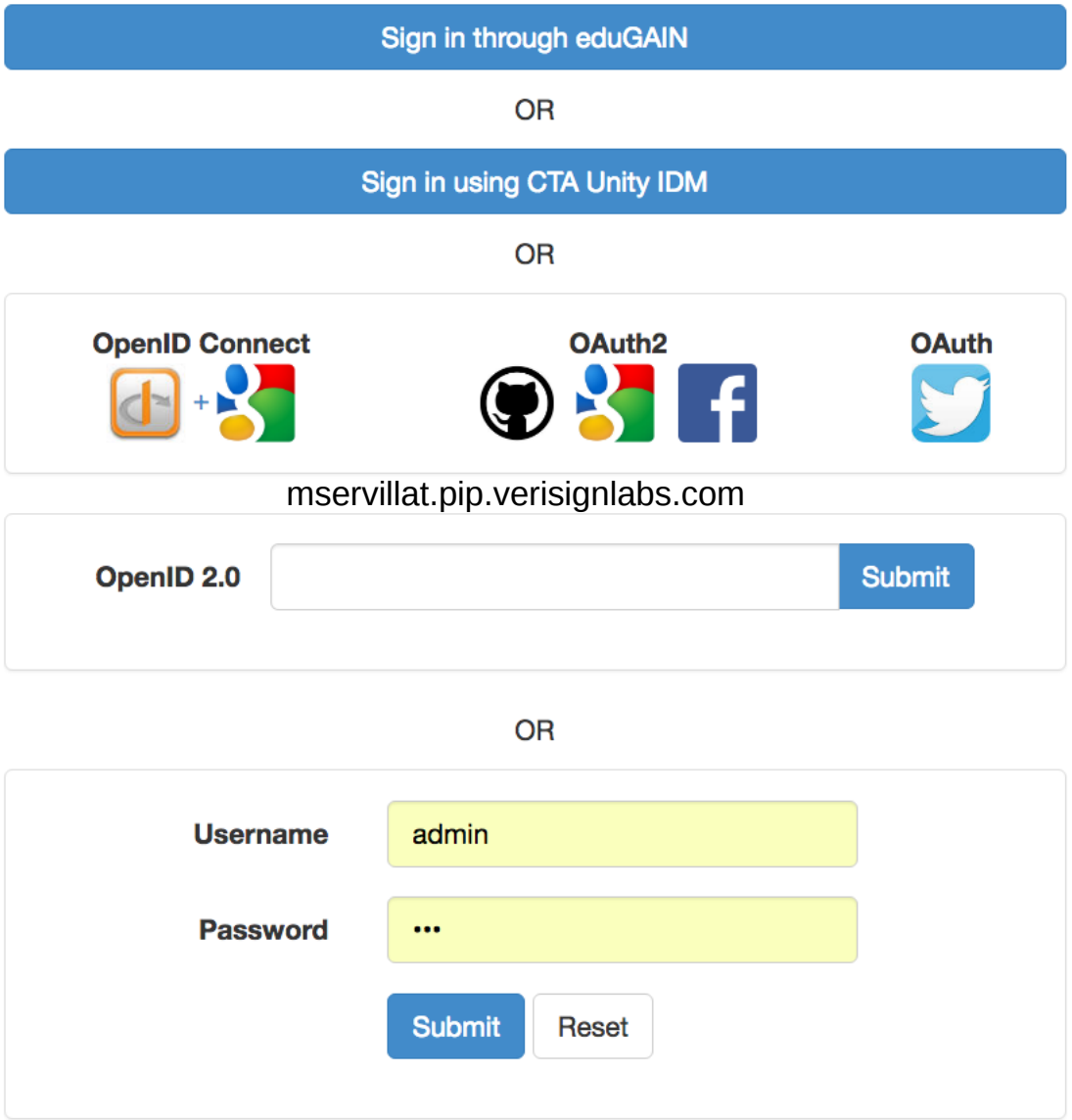
Submit | Reset

- Django, jQuery, BootStrap3
- Name resolver (Simbad through Sesame)
- Builds and Sends the ADQL query

ObsCore Search

proposal_id		Proposal ID
dataproduct_type	Nothing selected	Data product (file content) primary type
dataproduct_level	Nothing selected	DL0-5

Authentication & Authorization

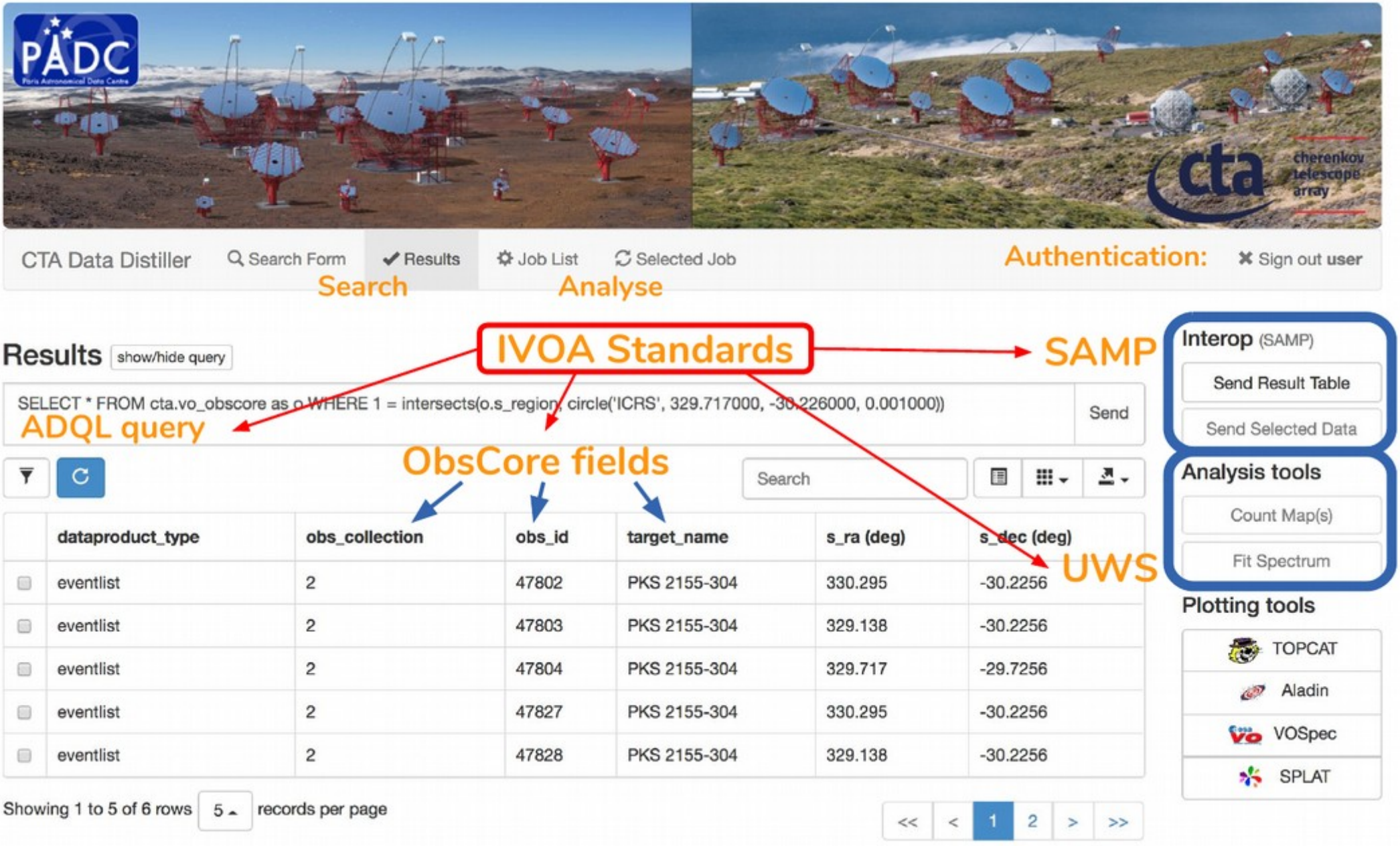


- **Shibboleth + Grouper**
  - EduGAIN federation
  - SAML2
- Unity IDM
  - Uses OpenID Connect
- **OpenID Connect**
  - Google as an IdP
- OAuth2
  - Github, Google, Facebook, ...
- OAuth
  - Twitter, ...
- OpenID 2.0 (deprecated)
- Local account

13

CTA Data Distiller

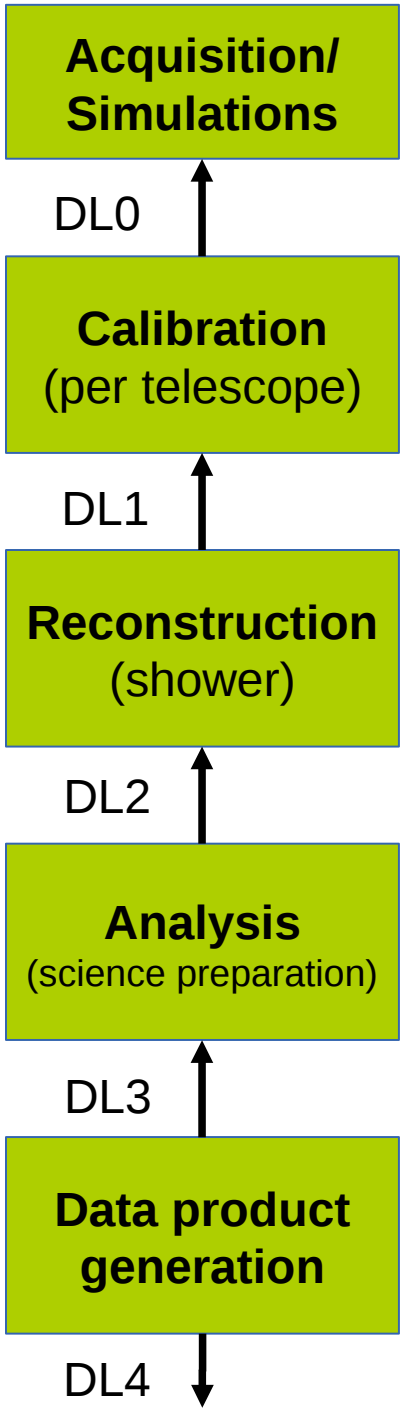
https://voparis-cta-test.obspm.fr



Pipeline requirements

- **Open** observatory
- A-USER-0110 : must ensure that data processing is **traceable** and **reproducible**
- **Inform** user on processing steps performed
- **Link to progenitor** to regenerate data (DL3 to DL4)

- Identify how a data product was produced
  ⇒ Provenance
- Identify what detailed options were used
  ⇒ Configuration



15

Data requirements

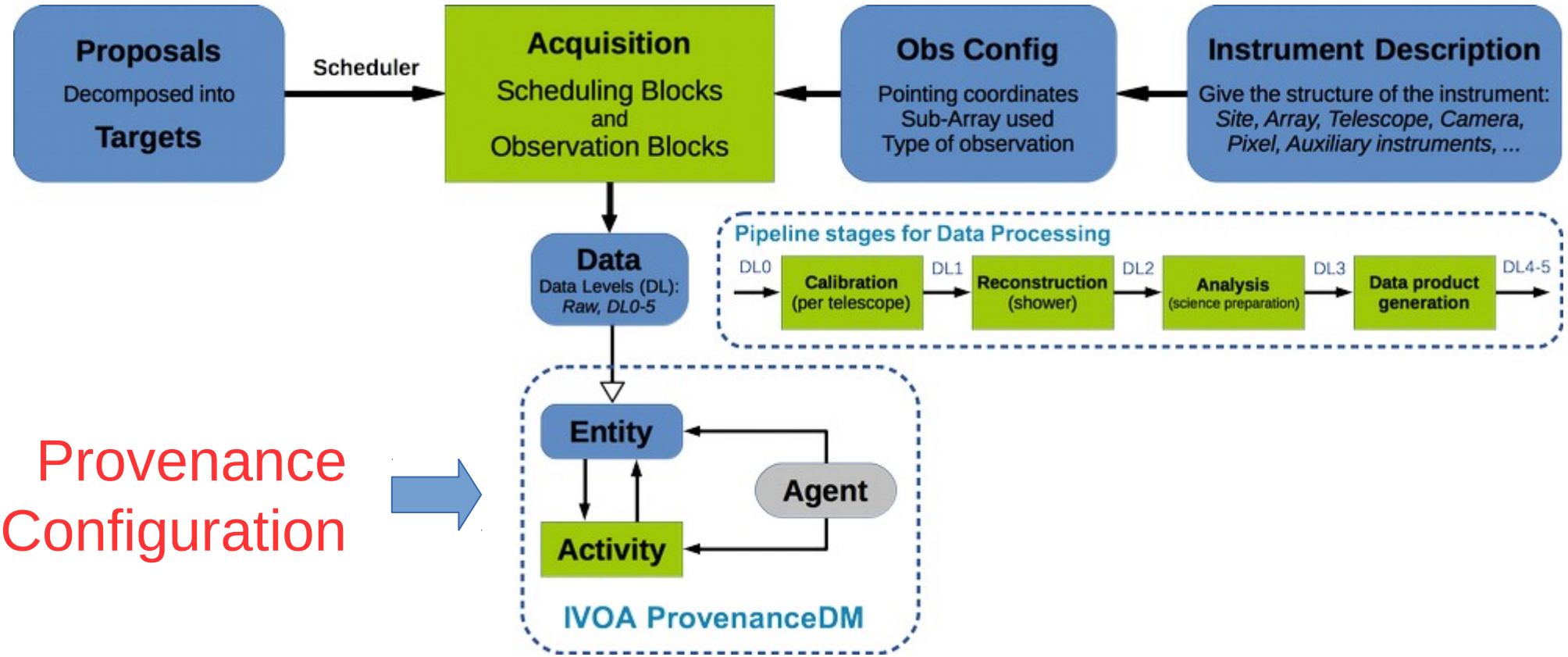
- C-DATA-MODEL-ALL-000050 :
  Data Model Processing history, software: The versions of the software release used for data taking, calibration and processing, etc of the data contained in a file will be stored as meta-data in the same file.
- C-DATA-MODEL-ALL-000052 :
  Data Model Processing history, characterization data: It will be possible to find the data which a file depends on, by using the metadata contained in the file itself. E.g. the previous data levels or the calibration data used to generate a file will be identifiable in this way.
- C-DATA-MODEL-ALL-000054 :
  Data Model Processing history, provenance: The provenance information of a file (creation center, creation date, etc) will be stored as metadata in the file.

16

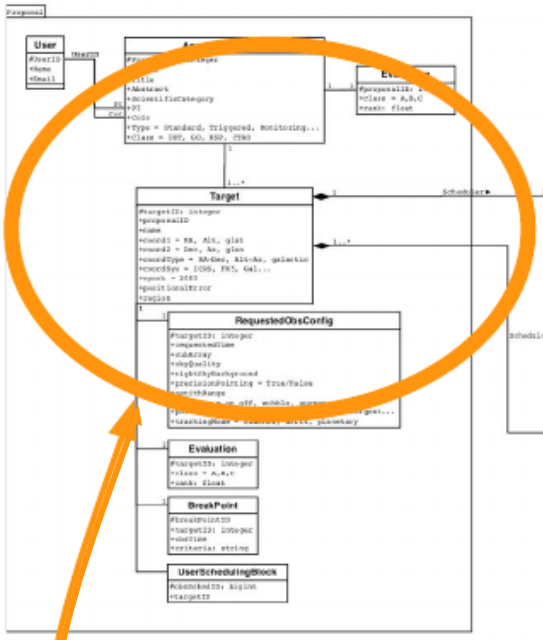
⇒ Covered by using the IVOA Provenance data model

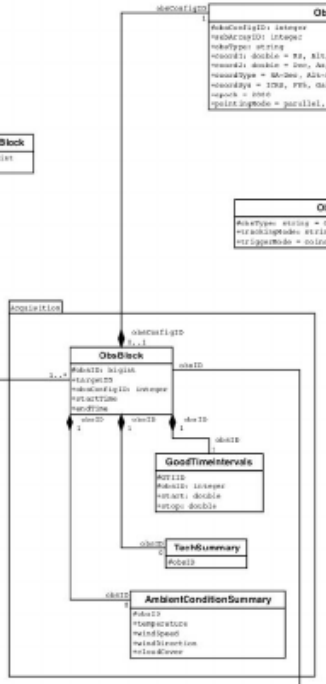
Master Configuration Data Model

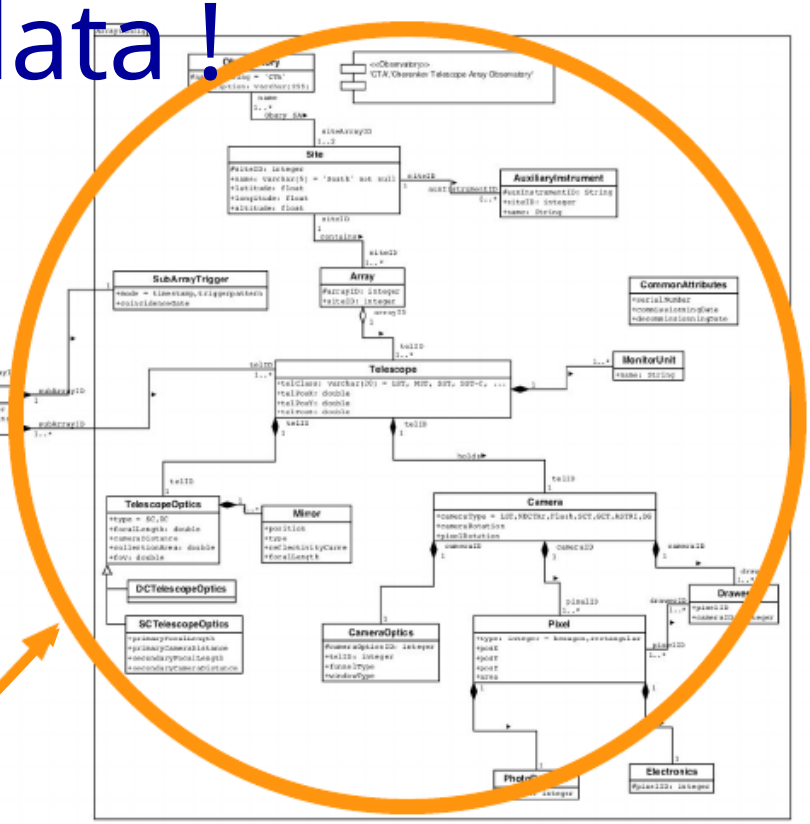
- Defines **structure** of services, content and context of data
- Can be seen as a **global interface**



All you need is metadata !

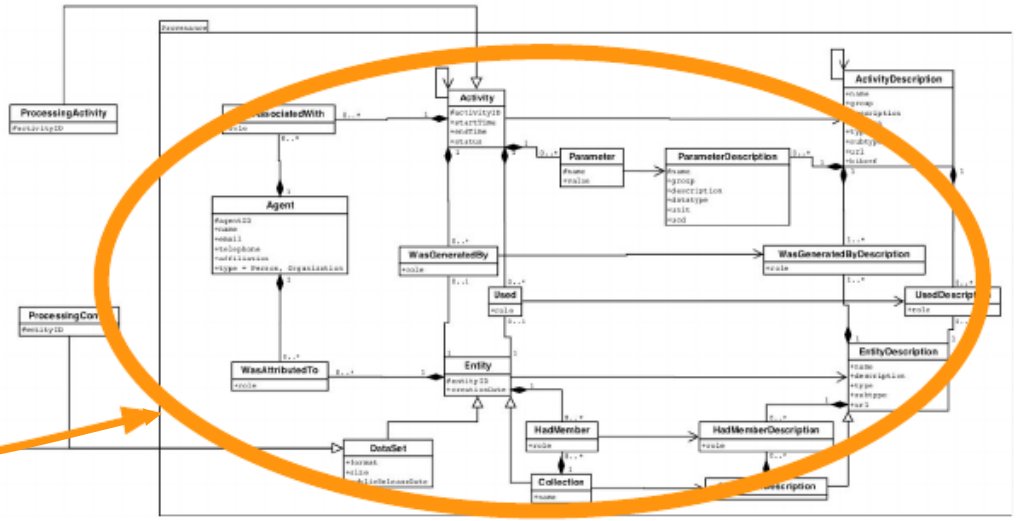






“Extended” Provenance

- Proposal
- Observing Mode
- Ambient Conditions
- Processing History

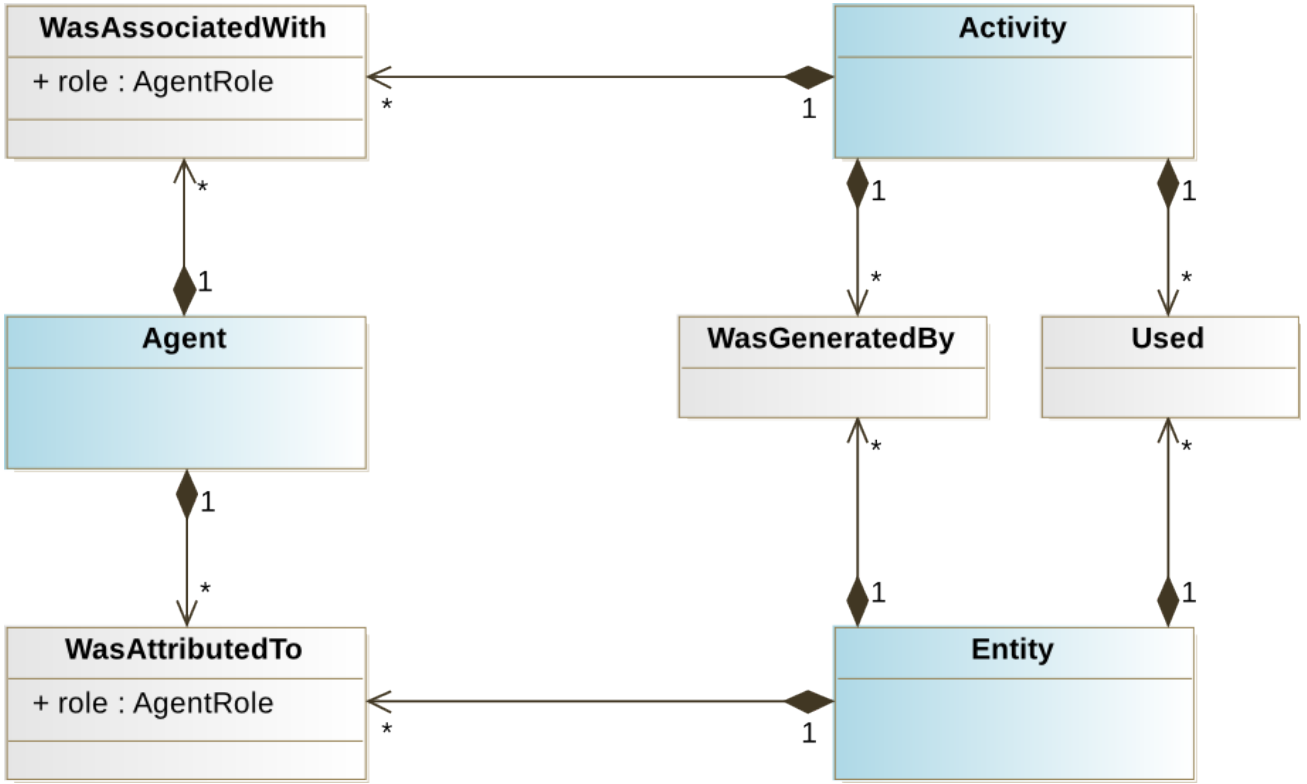


What kind of queries ?

Use case	Description
Cone Search	**Search data available for a given Target**
ObsCore search	**Search data available corresponding to ObsCore keywords** (target_name, time interval, ...), e.g.: • search data for a given **target** at a given **time** • search data in a given **region of the sky** • search data that contain events at **energy** higher than 50 TeV
ObsCore optional search	**Search data available corresponding to ObsCore optional keywords** (target_class, data_rights, ...), e.g.: • search **public** data for all **blazars** • search data for a given **proposal_id**
ObsConfig search	**Search data available corresponding to ObsConfig keywords** (sub_array_name, pointing_mode, obs_mode ...), e.g.: • search data that include the **Large Size Telescopes** (LSTs) • search data for a given target, that do not include the divergent **pointing mode**
Provenance search	**Search data available corresponding to Provenance keywords** (calib_version, creation_date ...), e.g.: • search data produced by a given **version of the pipeline** and for a given target • search data produced using a given **reconstruction method** • search data for a given target produced with **loose cuts**

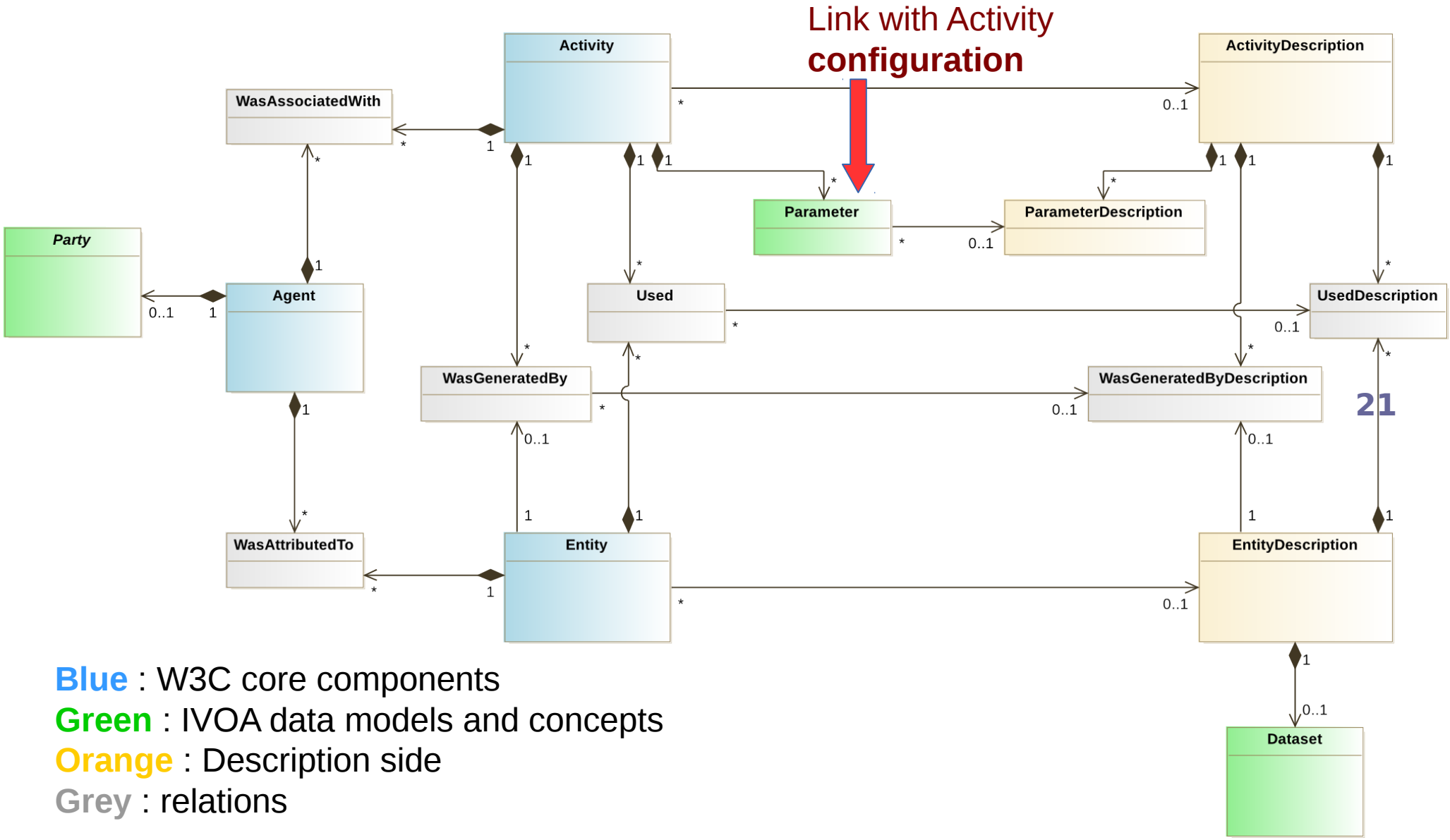
Provenance from W3C PROV

Provenance is “information about **entities**, **activities**, and **people** involved in producing a piece of data or thing, which can be used to form assessments about its **quality**, **reliability** or **trustworthiness**”.



W3C PROV Ontology : https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

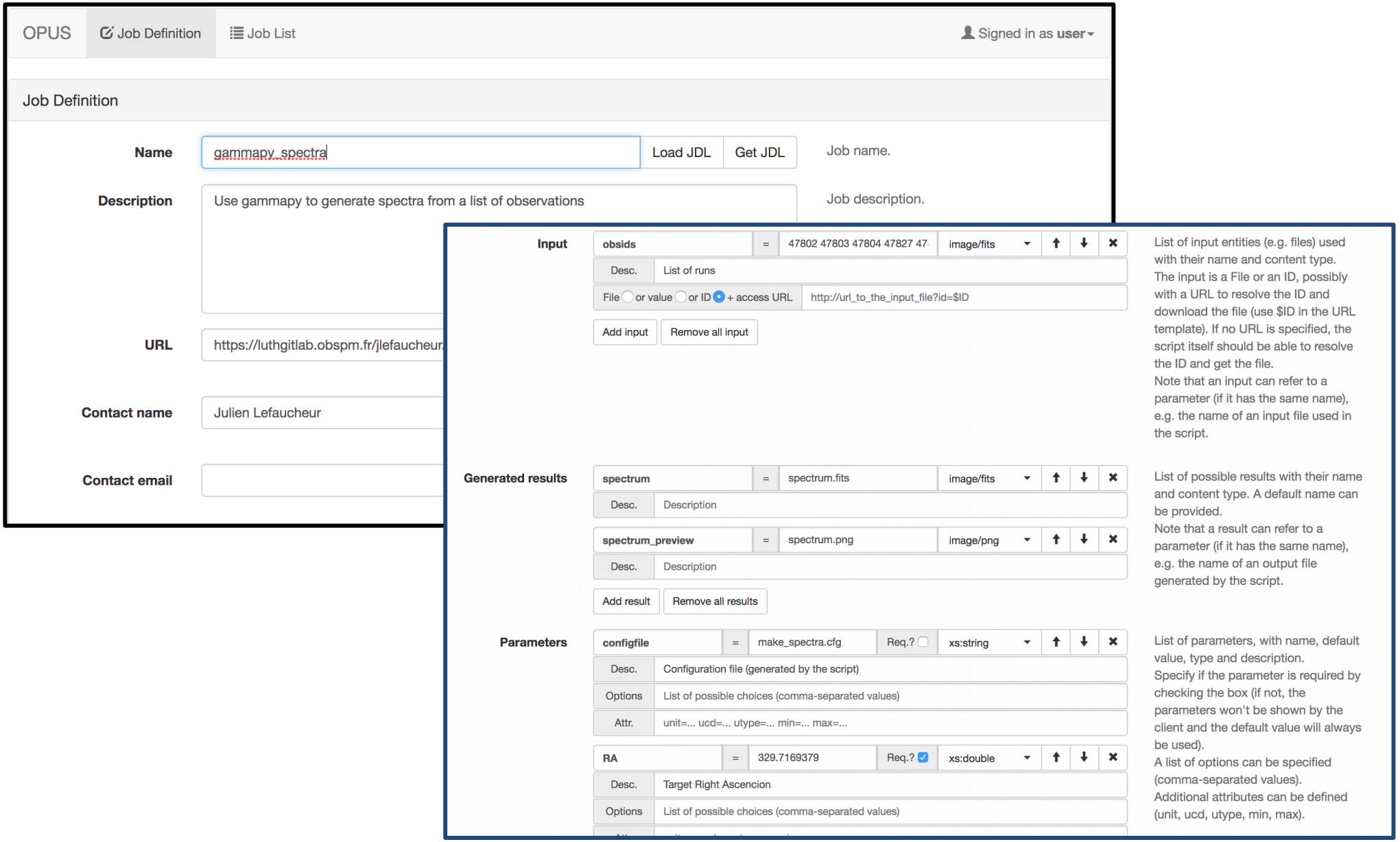
IVOA Provenance Data Model



Blue : W3C core components
Green : IVOA data models and concepts
Orange : Description side
Grey : relations

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

Description of a gammapy_spectra job



Web client working prototype

OPUS | Job Definition | Job List | Signed in as user

Job List for gammapy_spectra

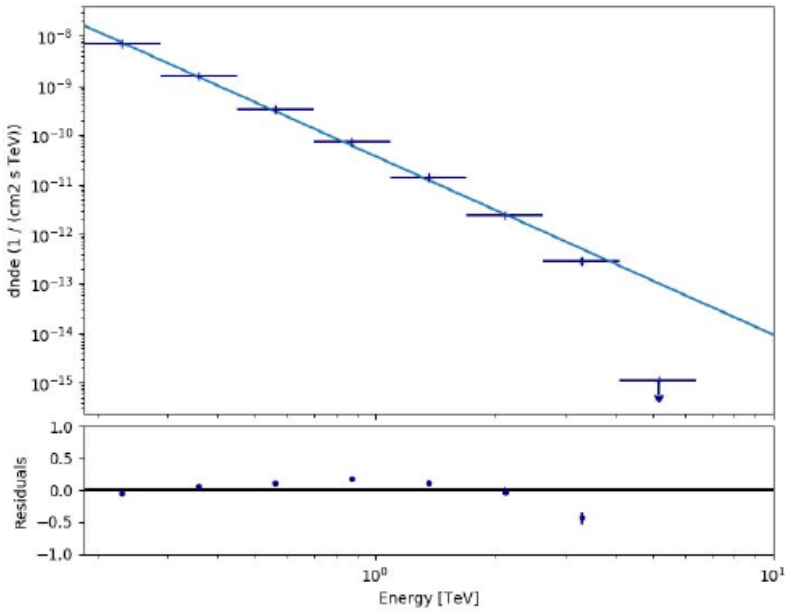
Refresh Job List | Create Test Job | Create New Job

Type	Start Time	Destruction Time	Phase	Details	Control
gammapy_spectra	2017-10-02 10:47:07	2017-11-01 10:47:05	COMPLETED	Properties | Parameters | Results	Start | Abort | Delete
gammapy_spectra		2017-11-01 10:47:03	PENDING	Prop	
gammapy_spectra	2017-09-29 15:07:52	2017-10-29 15:07:51	COMPLETED	Prop	
gammapy_spectra	2017-09-29 14:55:10	2017-10-29 14:55:09	ABORTED	Prop	
gammapy_spectra	2017-09-29 14:21:20	2017-10-29 14:21:19	COMPLETED	Prop	

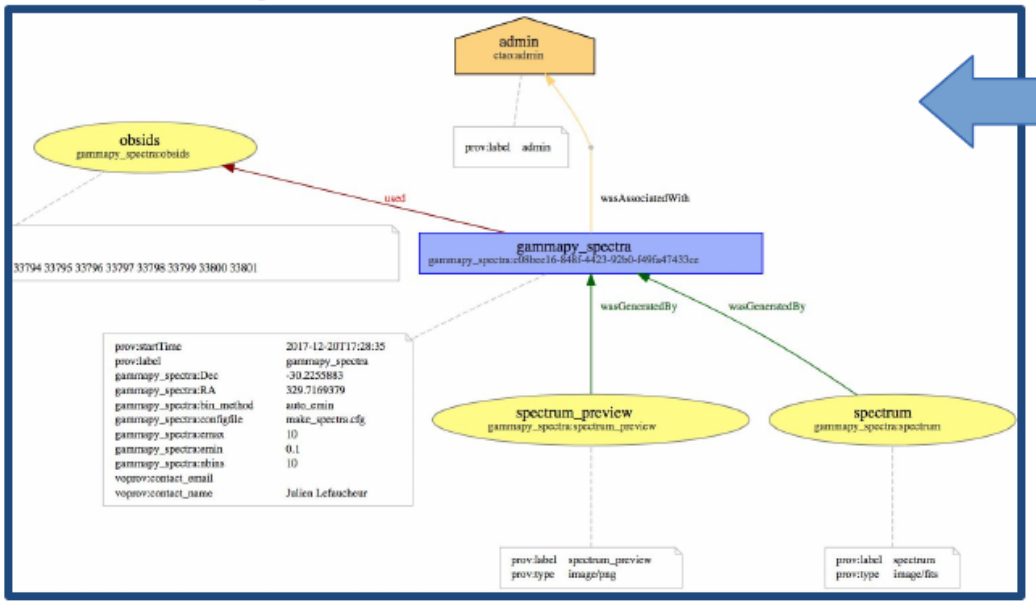
Job Results

spectrum: Download [image/fits]

spectrum_preview: Download [image/png]

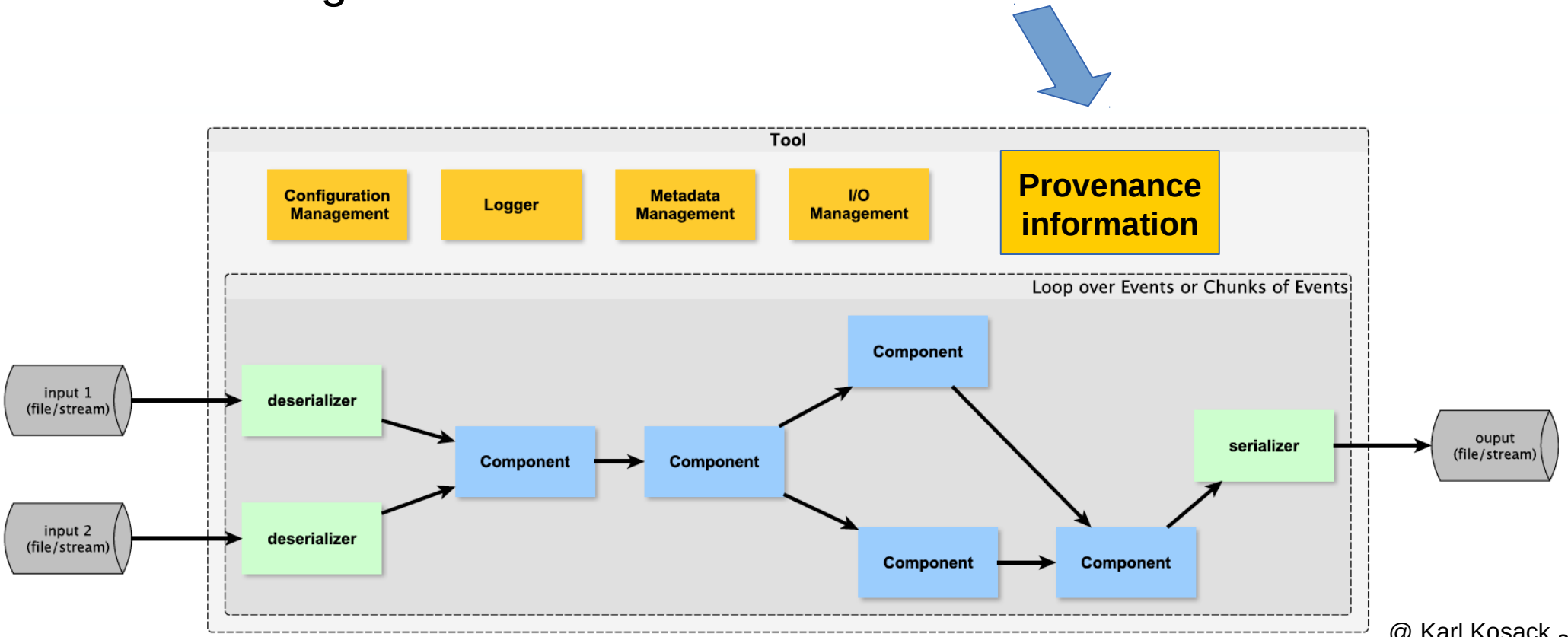


Tracking of Provenance informations



Provenance in the pipeline

- **Ctapipe**: a CTA data processing framework https://github.com/cta-observatory/ctapipe
- **Tool Python class** providing configuration, logger,metadata, I/O management… and **Provenance information**



Provenance class for ctapipe

```
from ctapipe.core import Provenance

prov = Provenance()
 # prov a singleton, so this gives you the same provenance class

prov.start_activitity("some_activity")

... # do things
prov.add_input_file("test.txt")
prov.add_output_file("out.txt")

prov.start_activity("some_sub_activity")

# do more things
prov.add_output_file("out2.txt")

prov.finish_activity() # finish some_activity
prov.finish_activity() # finish some_sub_activity
```

- Importance of **persistent identifiers**
- Also records **system configuration**, **state**, **software versions**

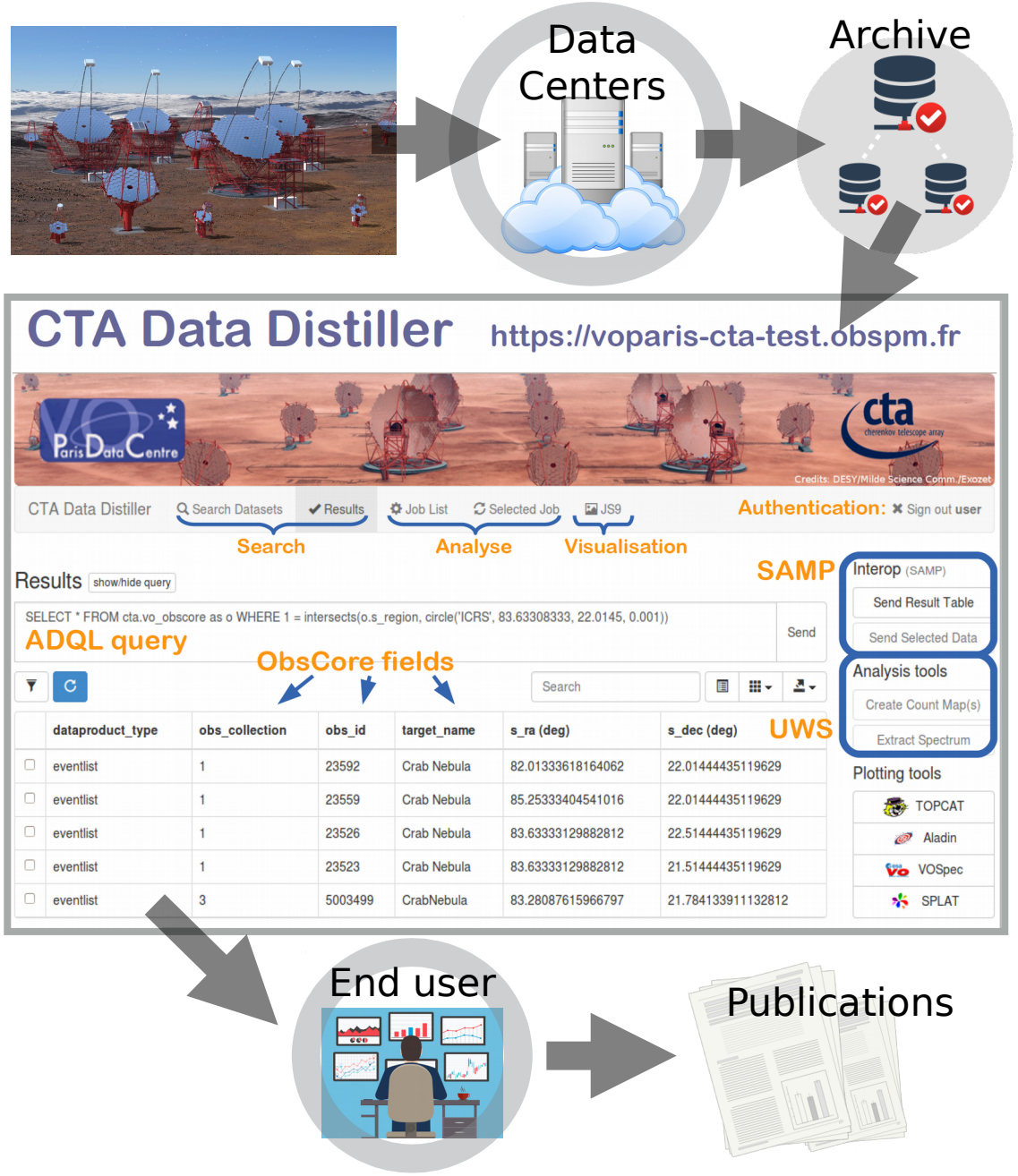
Behind the scene

- IVOA Provenance **data model** (CTA is a major use case)
- **Serialization** formats (W3C compatible, JSON/XML/…)
- Centralized Provenance **database** (prototypes available)
- **Access** services (ProvDAL and ProvTAP developed within the VO)

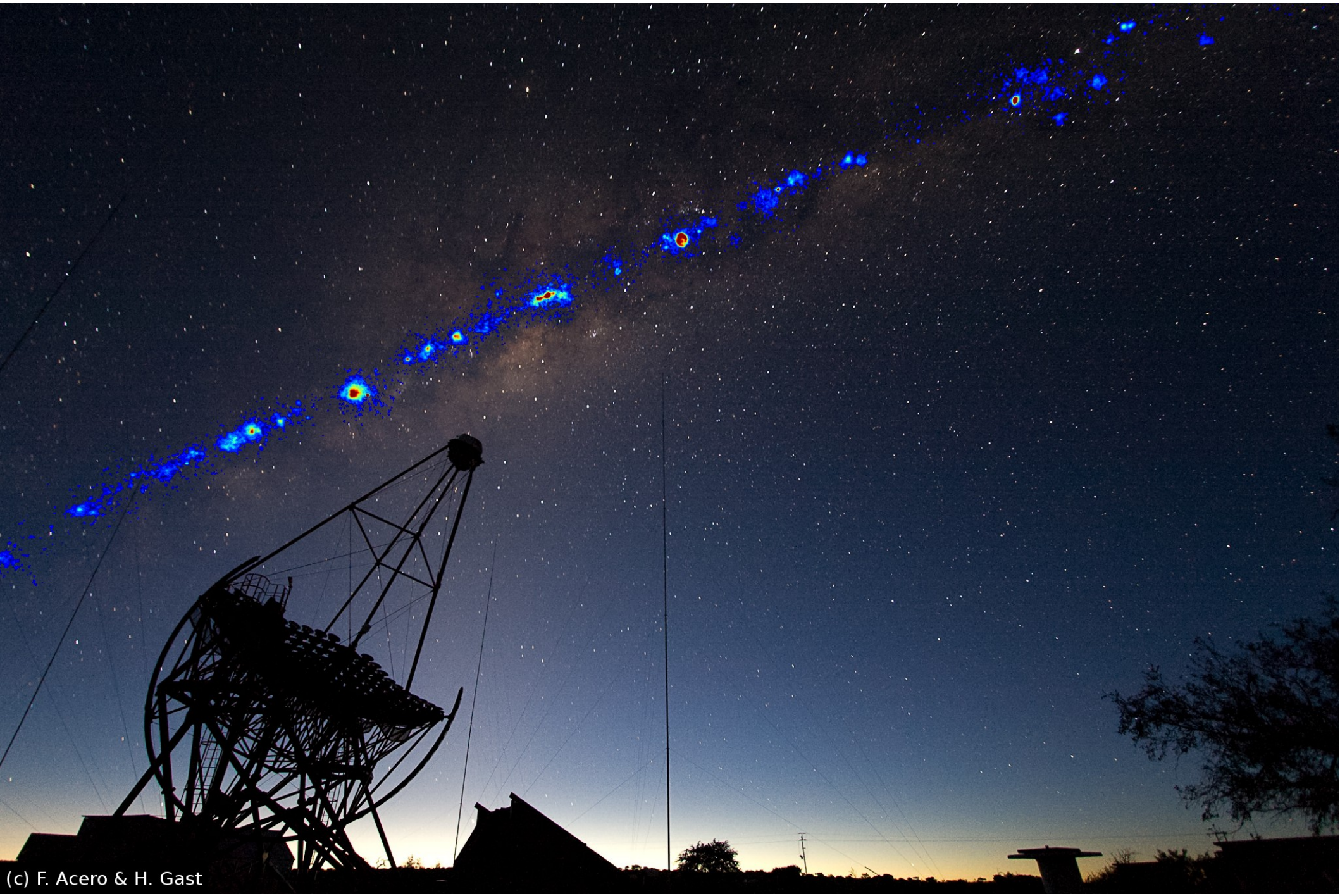
- **To be discussed:**
  - Definition of a dataset for CTA (events + IRF + … for DL3?)
  - Unique identifier for this dataset?
  - Data access queries
  - Provenance queries and views (e.g. what prov info for DL3?)

Science Archive and Science Gateway

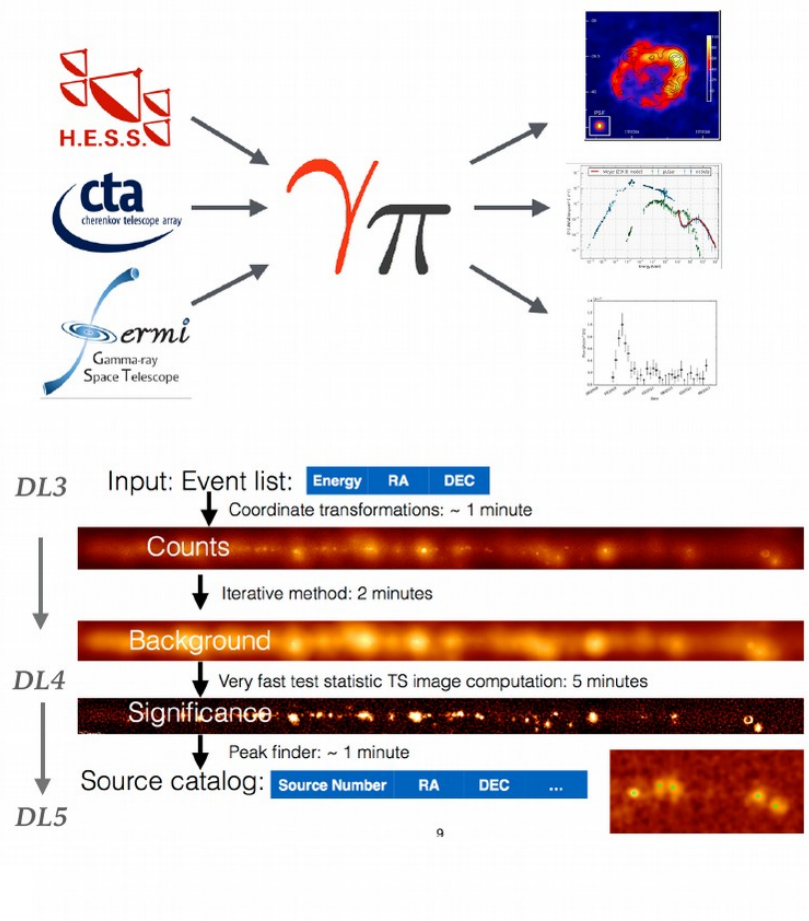
- Conception of a CTA Master Configuration **Data Model**
- Containing detailed **provenance** metadata stored in the **Archive**
- Compatibility with **Virtual Observatory** standards
- **Science Gateway** = collection of **interconnected** web services with common **Authentication/Authorization** system




© Fabio Acero


(c) F. Acero & H. Gast

Gammapy



- Python package
- Open development on Github
- Currently used for H.E.S.S., CTA preparation and Fermi-LAT
- Scope: science tools
  - DL3 (events, IRF,…)
  - DL4 (images, spectra,…)
  - DL5 (catalogs)

https://github.com/gammapy/gammapy

It’s a long way…

- *H.E.S.S, MAGIC & VERITAS have been operating independently for the last decade*
- *Variety of data formats and proprietary software, developed for each specific experiment.*
- *Field originally developed by particle scientists with a background biased towards particle physics rather than astronomy, and therefore with a different tradition regarding the data distribution formats.*

My data are too complicated for non expert users

My institute paid for building the experiment

May be there is more to get out of my original data

Want to know what is happening to my original data (keep an eye on science)

Open Archival Information System (OAIS)

Standard design for an **archive** to preserve information and make it available for a Designated Community (ISO 14721:2012)