



Leibniz-Institut für
Astrophysik Potsdam

A Provenance Data Model for astronomy

ADASS XXVI, 19th October 2016, Trieste

Kristin Riebe

François Bonnarel

Mireille Louys

Florian Rothmaier

Michèle Sanguillon

Mathieu Servillat

IVOA Data Model Working Group



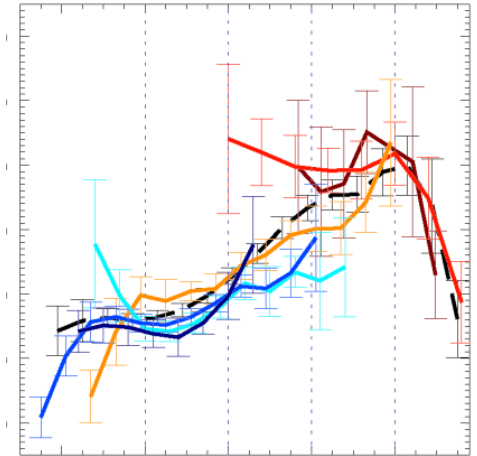
What is provenance?

- In general: tracking the history, origin of something:
 - art
 - food industry
 - information (data vis) on news webpage
 - scientific data!



What is provenance?

- In general: tracking the history, origin of something:
 - art
 - food industry
 - information (data vis) on news webpage
 - scientific data!
- In astronomy: explain how data sets were produced:
 - Who created the data?
 - Which algorithm was used to produce it?
 - Which steps were undertaken to process the image?
 - Can I get access to the original, uncalibrated files from the observation?



Goals for IVOA Provenance Data Model

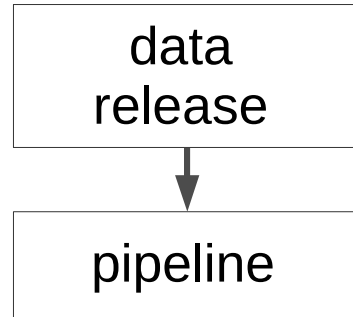
- For a given data set, provenance should help to ...
 - Discover steps of production
Aid in reprocessing: Which processing steps have been done already?
 - Give attribution
Who was involved in the project? Who can I ask about these data?
 - Allow to assess the quality of the data
Is the dataset suited for my research?
 - Aid in debugging
Find possible error sources, e.g. check version of processing software, ambient conditions, telescope configuration, parameter settings, ...
 - Search in structured provenance metadata
Includes „forward tracking“: which datasets were produced with the same pipeline version, follow scientific productivity of instruments/telescopes or software usage

Example in astronomy

data
release

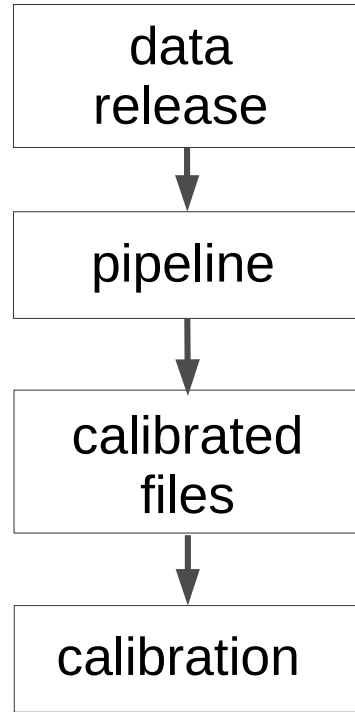
- Where is the data coming from?

Example in astronomy



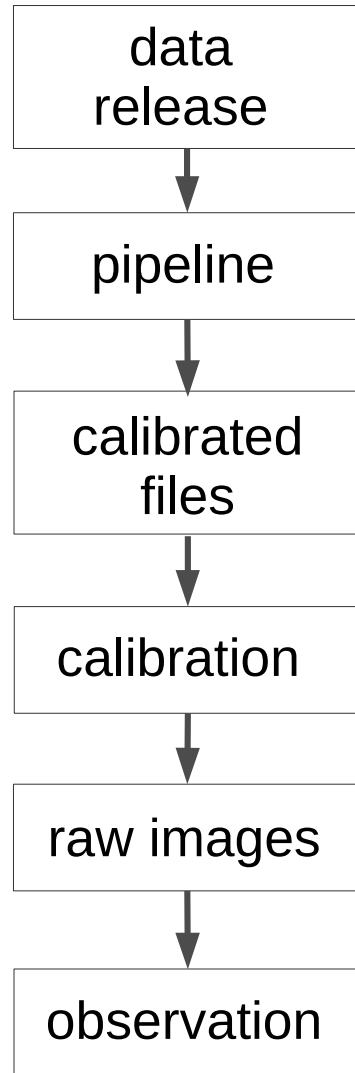
- Where is the data coming from?
- What were the input files for the pipeline?

Example in astronomy



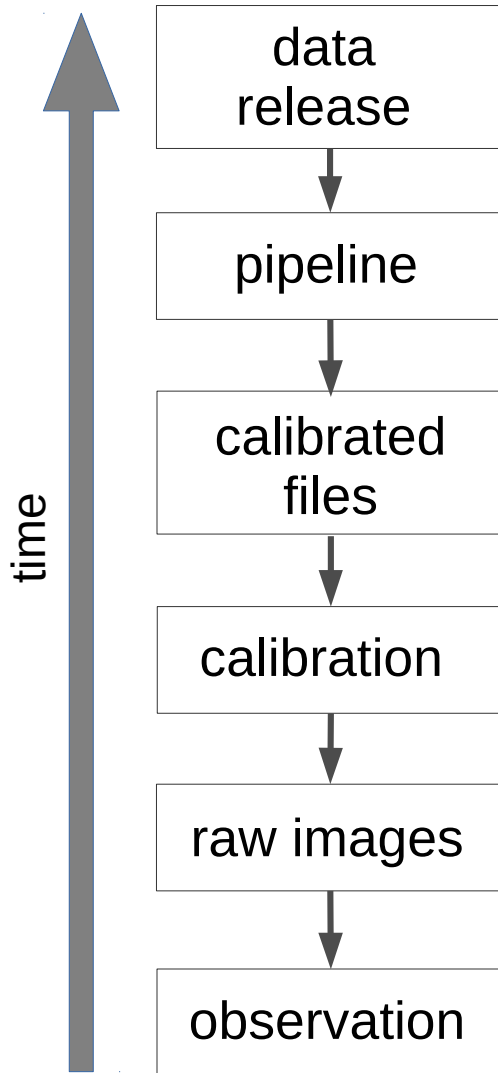
- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?

Example in astronomy



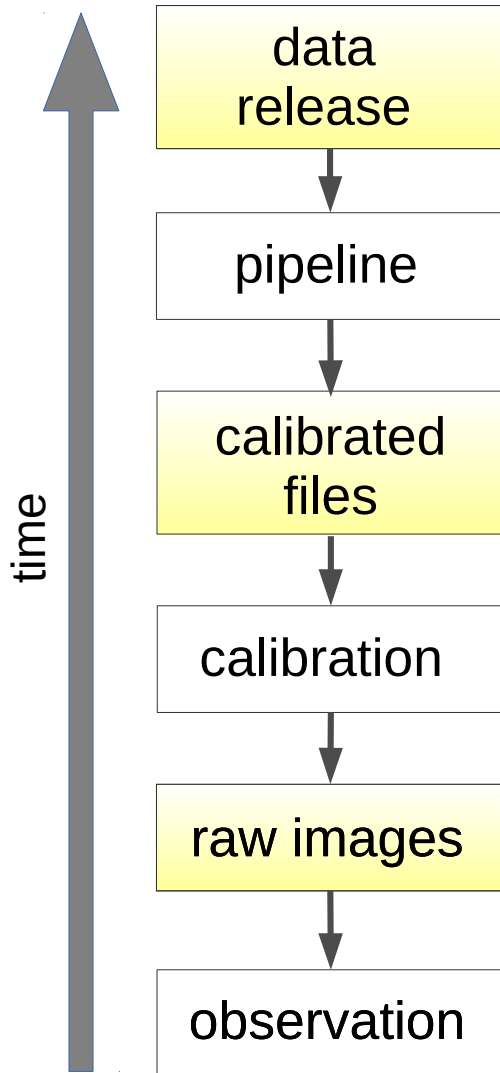
- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?
- Can I get the raw images?
- Were there perfect seeing conditions during the observation?

Example in astronomy

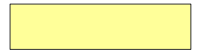


- Where is the data coming from?
 - What were the input files for the pipeline?
 - Have calibrated files been used for the pipeline?
 - How were they calibrated?
 - Can I get the raw images?
 - Were there perfect seeing conditions during the observation?
- => Track data back in time

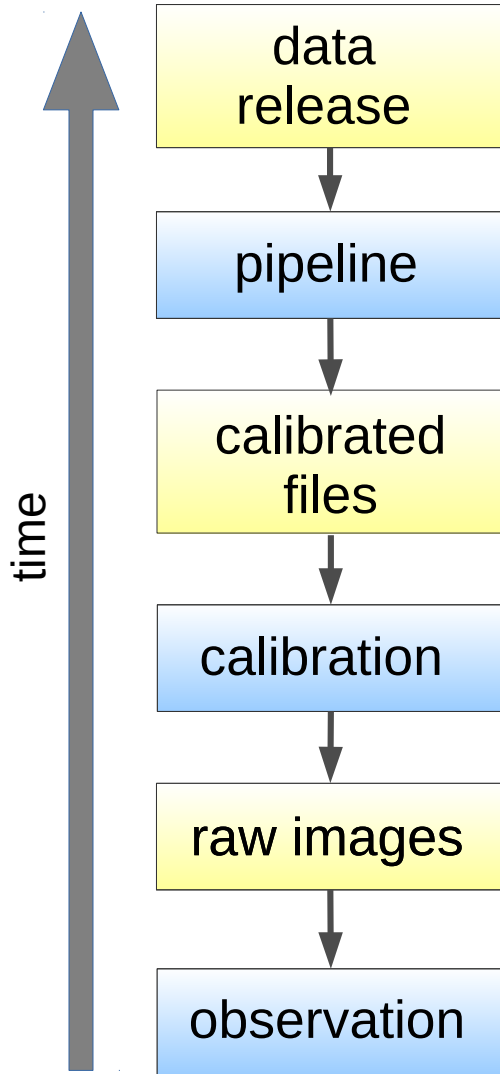
Example in astronomy



- identify data entities

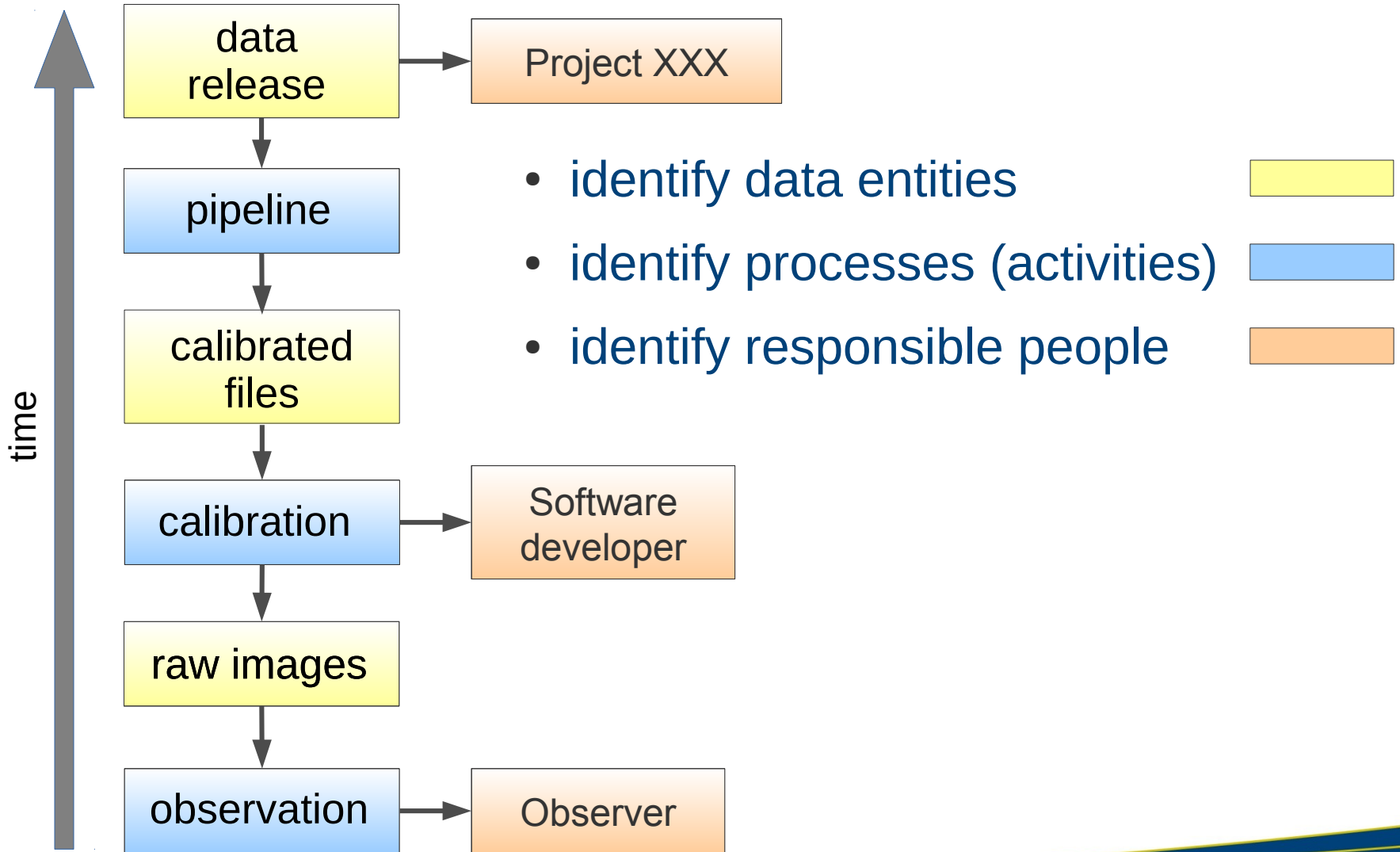


Example in astronomy

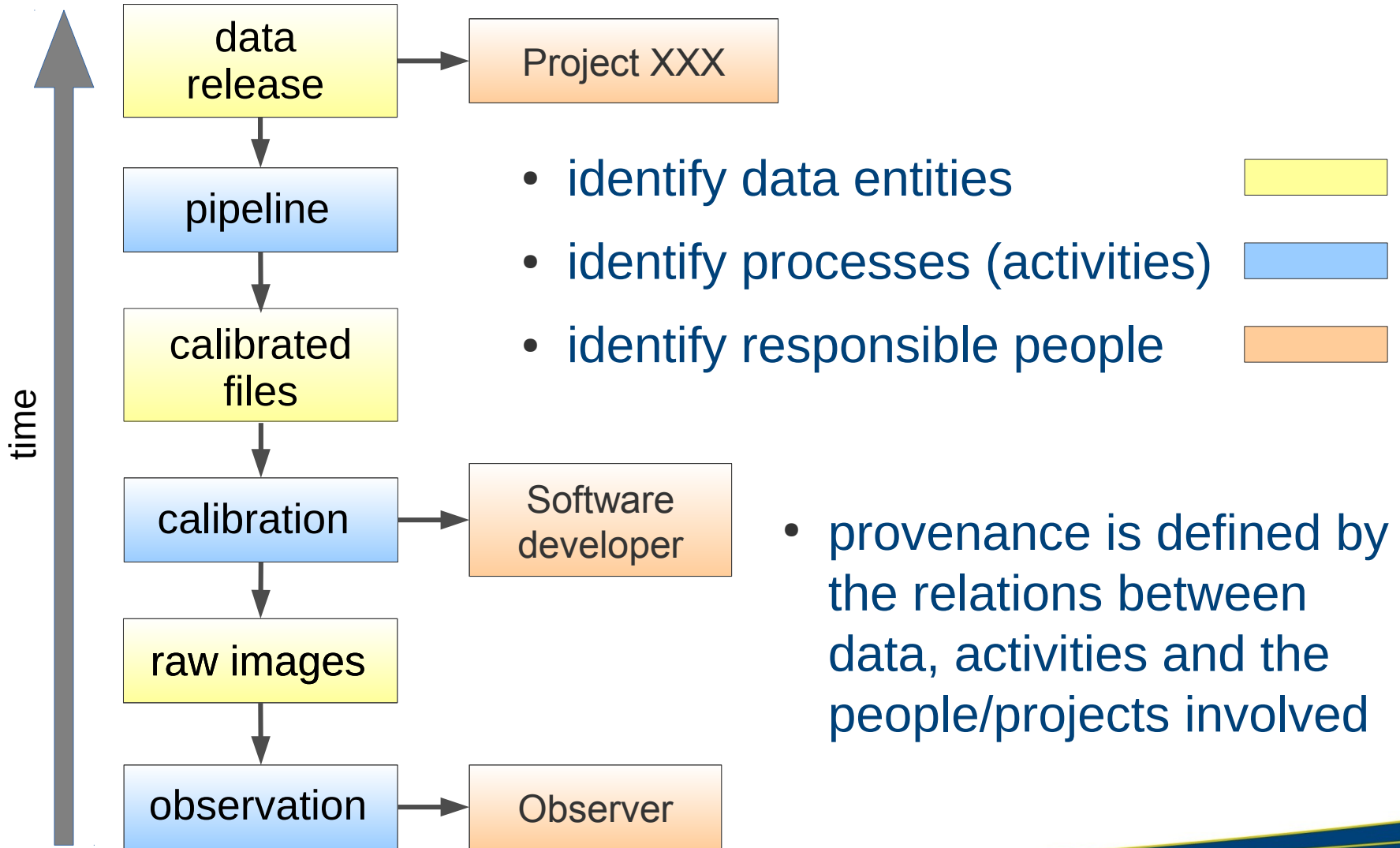


- identify data entities 
- identify processes (activities) 

Example in astronomy

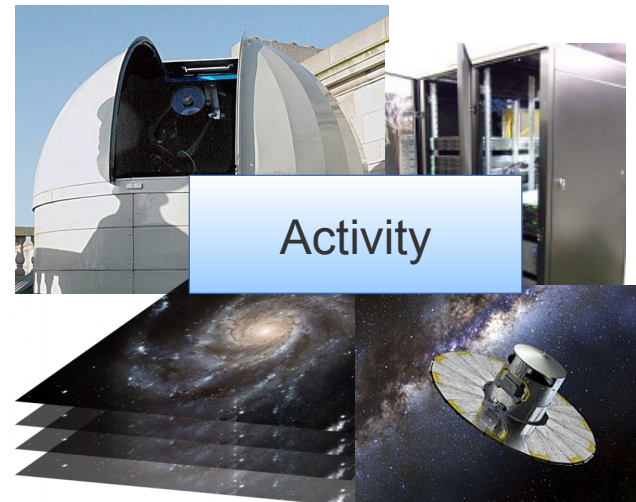
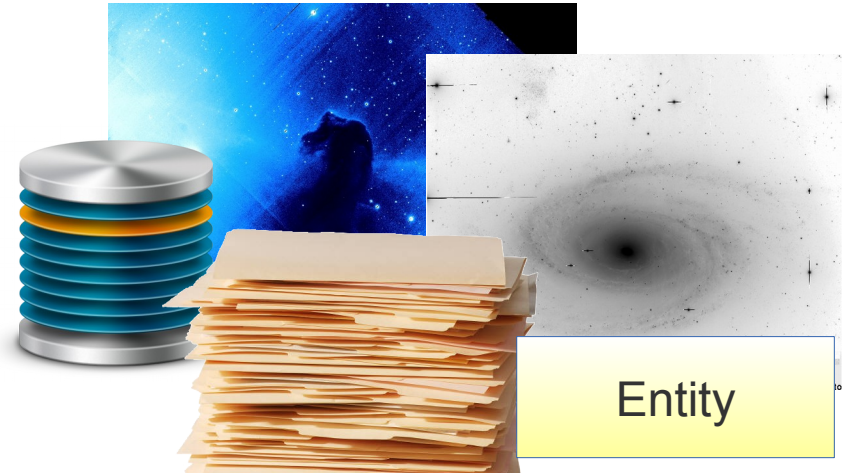


Example in astronomy



Examples for core objects

- **Entities (datasets):**
images, catalogs,
database tables, spectra,
log files, parameters, ...
- **Activities:**
observations;
processing steps like bias subtraction,
image stacking, continuum fit, object
extraction; simulations, ...
- **Persons/Organizations:**
creator, publisher, developer, ...



Provenance DM core classes

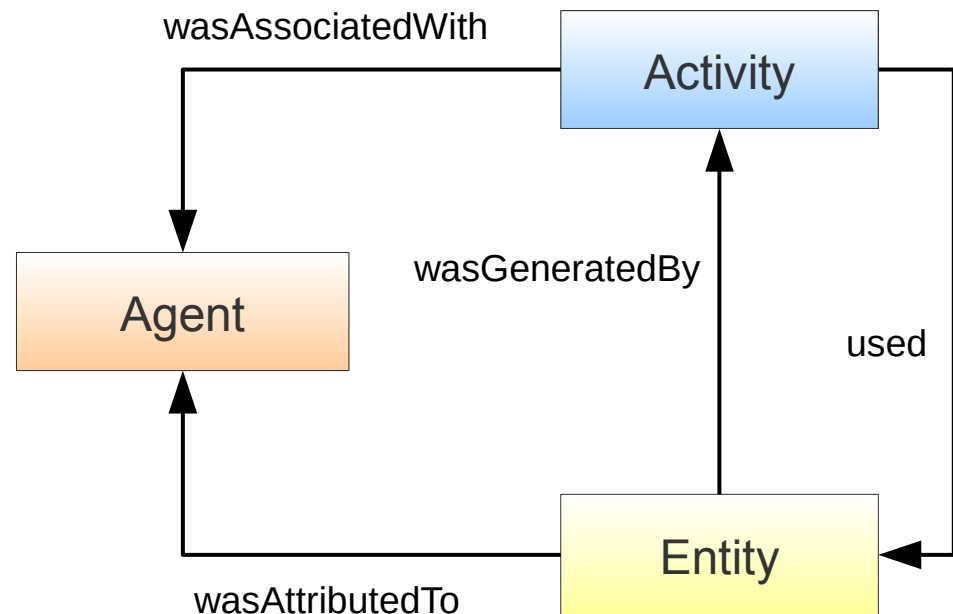
- same core classes as in W3C ProvDM model:
 - <http://www.w3.org/TR/prov-dm/>, published 2013

- 3 core classes:

- Activity
- Entity
- Agent

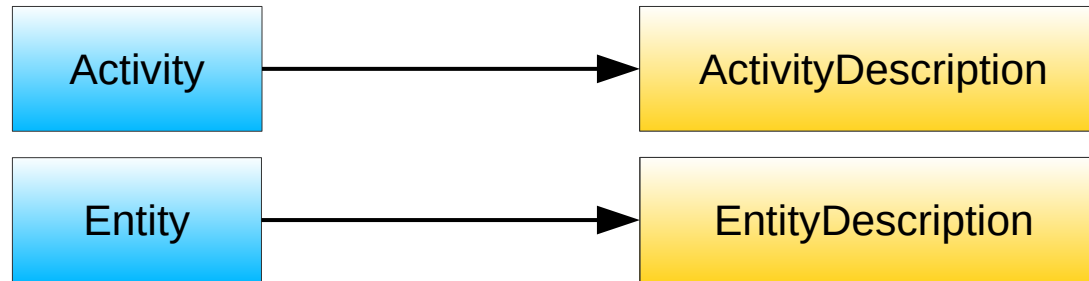
- core relations:

- used
- wasGeneratedBy
- wasAttributedTo
- wasAssociatedWith



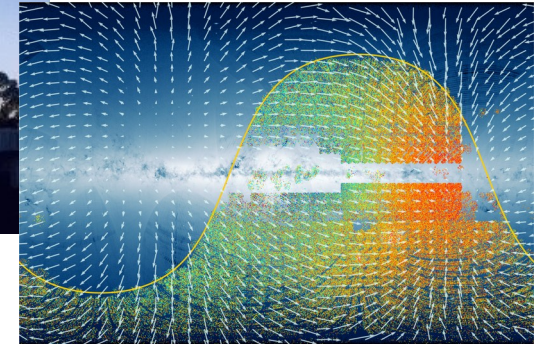
Extending the core

- in astronomy: know most common processes
- introduce new “description” classes for common processes and datatypes:

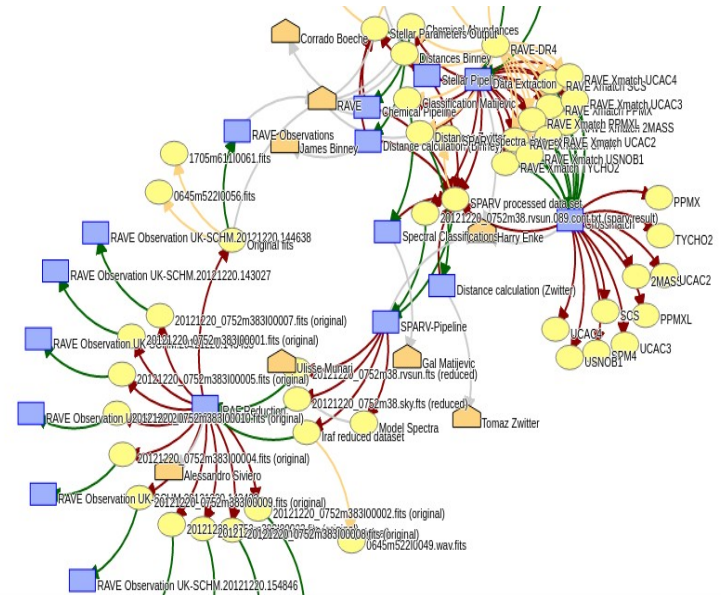


- connection to similar structures in other data models:
 - Activity => **Experiment** in Simulation Data Model
 - ActivityDescription => **Protocol** in Simulation Data Model
 - EntityDescription => **Dataset** in Dataset Data Model

Use case: RAVE




- multi-fibre spectroscopic survey
- radial velocities + derived stellar properties for ~ half million stars
- use provenance to track e.g.
 - Who was responsible for determining the log g values in DR5?
 - Which fibre observed the spectrum for star xyz?
 - Study selection effects using information on intended and actually observed stellar sample




see javacript example at <https://escience.aip.de/prov/graphs/example.html>

Use case: CTA


- see Poster by **Mathieu Servillat: P5.5 (upstairs)**
- must ensure that data processing can be traced and reproduced
- essential to inform users about processing steps



cherenkov telescope array
an observatory for ground-based gamma-ray astronomy



Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477



Structuring metadata for the Cherenkov Telescope Array

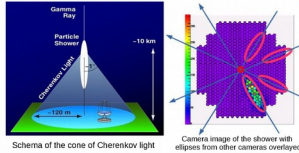
Mathieu Servillat¹, Catherine Boisson¹, Julien Lefaucheur¹, Johan Bregeon², Michèle Sanguillon² and Jose-Luis Contreras³ for the CTA Consortium⁴

¹Laboratoire Univers et Théories, Observatoire de Paris / CNRS / PSL, Meudon, France
²Laboratoire Univers et Particules de Montpellier, France
³Universidad Complutense de Madrid, Spain
⁴See http://bit.do/cta_consortium for full author & affiliation list

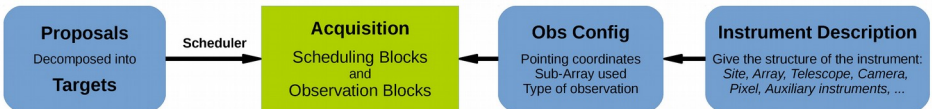
ABSTRACT
 The landscape of ground-based gamma-ray astronomy is drastically changing with the perspective of the **Cherenkov Telescope Array (CTA)** composed of more than 100 Cherenkov telescopes. For the first time in this energy domain, CTA will be operated as an observatory **open** to the astronomy community. In this context, a structured **high level data model** is being developed to describe a CTA observation. The data model includes different classes of metadata on the project definition, the configuration of the instrument, the ambient conditions, the data acquisition and the data processing. This last part relies on the **Provenance Data Model** developed within the International Virtual Observatory Alliance (IVOA), for which CTA is one of the main use cases. The CTA data model should also be compatible with the Virtual Observatory (VO) for data diffusion. We have thus developed a web-based data diffusion prototype to test this requirement and ensure the compliance.

Objectives
 The high level data model diagram aims at defining **global terms** and their **relations**, in order to provide the complete description of a CTA data product. This data model or part of it is relevant to **various CTA working groups**: Proposal Handling, Array Control, Pipeline, Data Diffusion and Hardware Developments. It serves as a **global interface**.

Cherenkov Astronomy
 The **Imaging Atmospheric Cherenkov Technique (IACT)** is a method to detect very high energy gamma-ray photons in the **50 GeV to 50 TeV** range. It works by imaging the very short flash of Cherenkov radiation generated by the cascade of relativistic charged particles (**shower**) produced when a very high-energy gamma-ray strikes the atmosphere.



Schema of the cone of Cherenkov light
 Camera image of the shower with ellipses from other cameras overlaid




Proposals (Decomposed into Targets) → Scheduler → Acquisition (Scheduling and Observation Blocks) ← Obs Config (Pointing coordinates, Sub-Array used, Type of observation) ← Instrument Description (Give the structure of the instrument: Site, Array, Telescope, Camera, Pixel, Auxiliary instruments, ...)

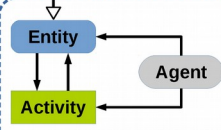
CTA High Level Data Model
 We show here the global structure only, without details on classes and attributes:

- **Proposals** are decomposed into **Targets** with their **requirements** (observing and pointing modes, ...), and **constraints** (e.g. night sky background, ...)
- The **Scheduler** then creates an observation program composed of blocks: **Scheduling Blocks** (sequence of observations planned for a given Target), made of **Observation Blocks** (effective start and stop times of acquisition with a given configuration)
- The **Obs Config** defines the coordinates, the **SubArray** (group of telescopes used), the **type** of observation, the strategy and the observing, pointing and trigger modes
- The **Instrument Description** is a separate database that contains the complete instrument description and its modifications
- **Raw Data** is produced during **Acquisition** and processed to higher **Data Levels**

Pipeline stages for data processing



VO Diffusion for CTA
 One of the goal of the High Level Data Model is to make **CTA data products available and discoverable** through the Virtual Observatory (VO). For example, the attributes contained in this data model can be mapped to the generic **IVOA ObsCore data model**, and exposed using the **IVOA Table Access Protocol (TAP)**. This provides an **ObsTAP service** for the CTA Archive. An **online prototype** has been developed to test the data model and adapt the VO protocols to Cherenkov Astronomy: <https://voparis-cta-test.obspm.fr>

Entity-Activity-Agent (EAA) Model

 The **tracking of processing activities** will be done using the IVOA Provenance Data Model, based on the **W3C PROV** ontology (Entity-Activity-Agent relations). This data model and its access layer are currently in development (see talks **O10.4** and **I10.1**).

Next developments
 This work is still **preliminary**. Most of the content of the data model is now defined but it still requires iterations with involved working groups to be completed.

Acknowledgements: ASTERICS (<http://www.asterics2020.eu/>) is a project supported by the European Commission Framework Programme Horizon 2020 Research and Innovation action under grant agreement n. 653477. Additional funding was provided by the INSU (Action Spécifique Observatoire Virtuel, ASOV), the Action Fédératrice CTA at the Observatoire de Paris and the Paris Astronomical Data Centre.

What's your use case?

- Would you benefit from a standardized solution to expose your provenance metadata?
=> contact us!
- How do you currently keep track of the data history?
- Which metadata would you need most?

What's your use case?

- Would you benefit from a standardized solution to expose your provenance metadata?
=> contact us!
- How do you currently keep track of the data history?
- Which metadata would you need most?

Talk to us and join discussions in
IVOA data model working group!