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

GERMAN ASTROPHYSICAL GAVO VIRTUAL OBSERVATORY

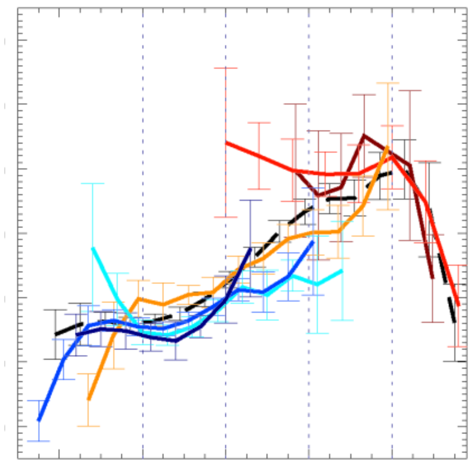
SPONSORED BY THE Federal Ministry of Education and Research

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

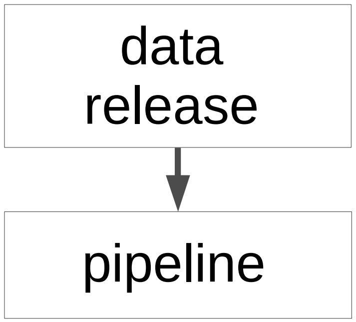
    Includes „forward tracking“: which datasets were produced with the same pipeline version, follow scientific productivity of instruments/telescopes or software usage

Example in astronomy

data
release

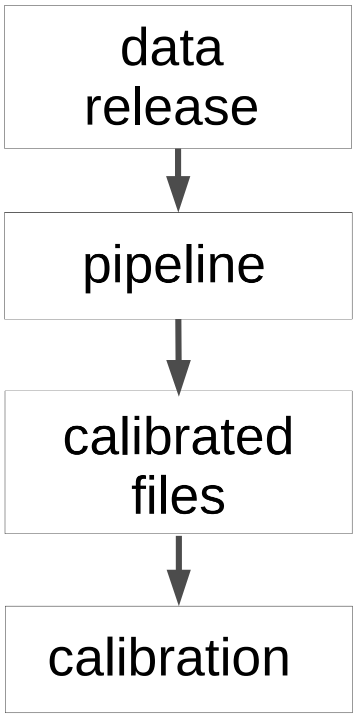
- Where is the data coming from?

Example in astronomy



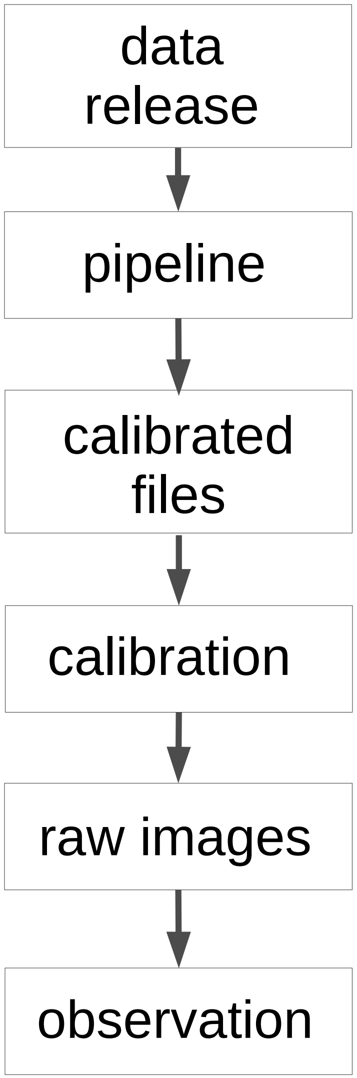
- Where is the data coming from?
- What were the input files for the pipeline?

Example in astronomy



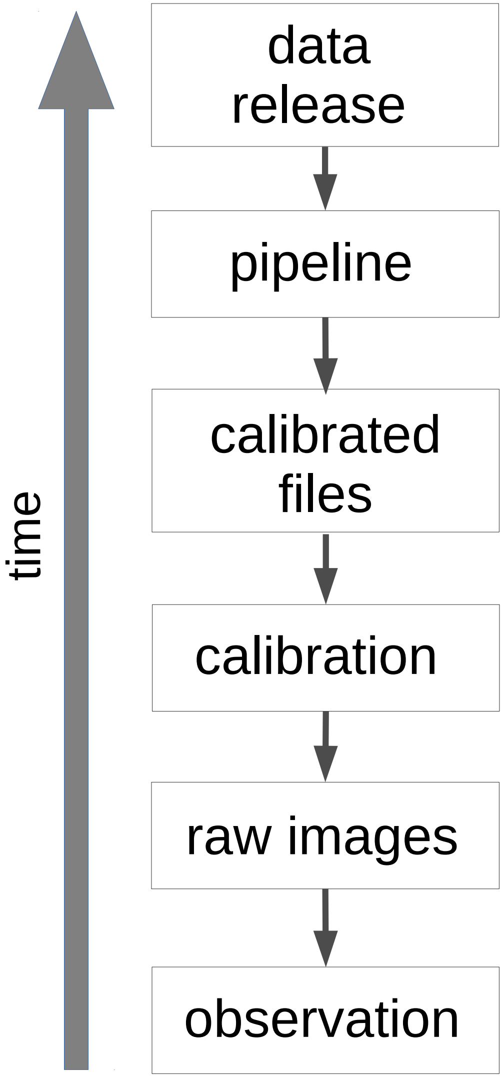
- Where is the data coming from?
- What were the input files for the pipeline?
- Have calibrated files been used for the pipeline?
- How were they calibrated?

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?

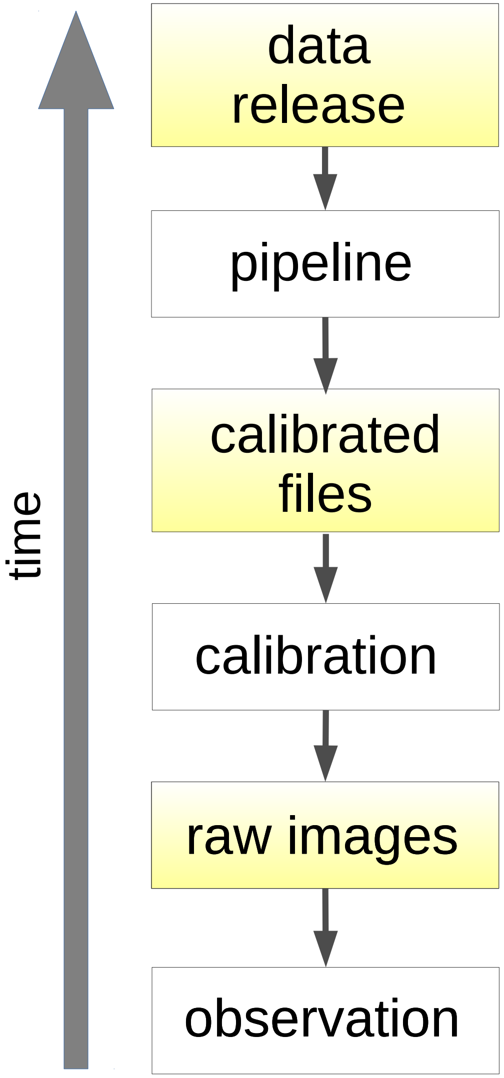
Example in astronomy



- 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?

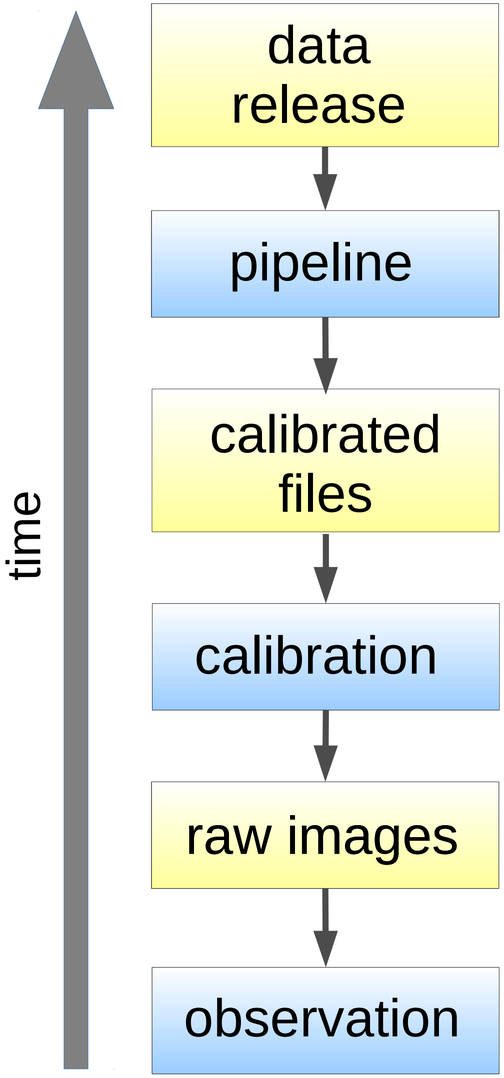
=> Track data back in time

Example in astronomy



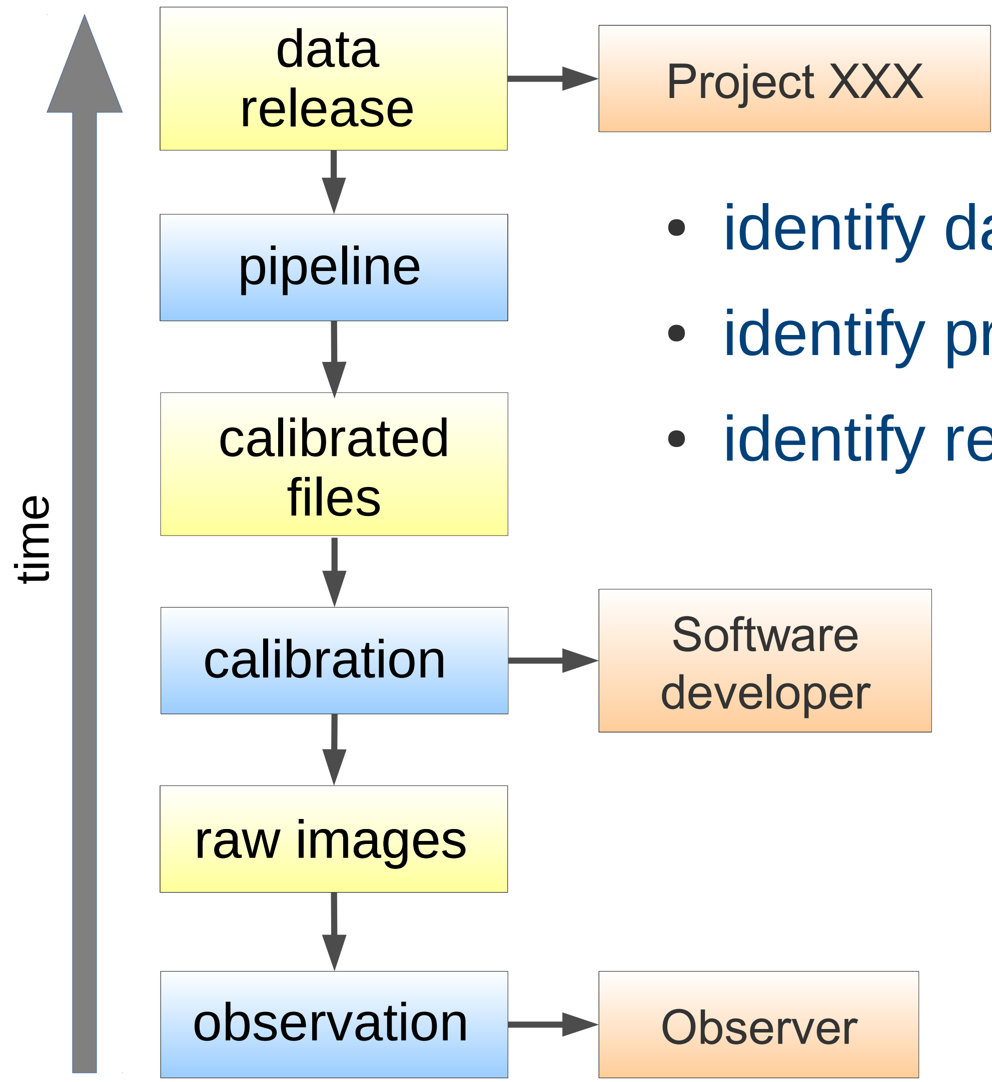
- identify data entities

Example in astronomy



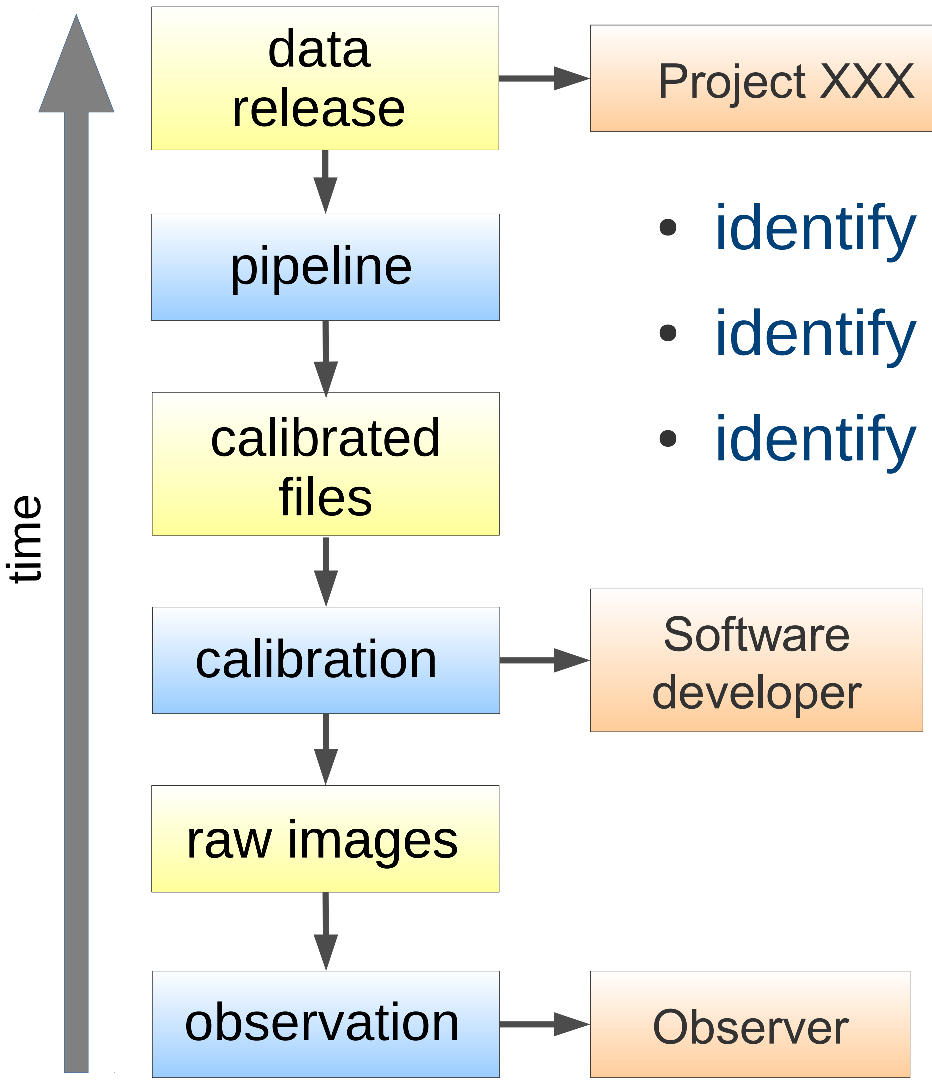
- identify data entities
- identify processes (activities)

Example in astronomy



- identify data entities
- identify processes (activities)
- identify responsible people

Example in astronomy

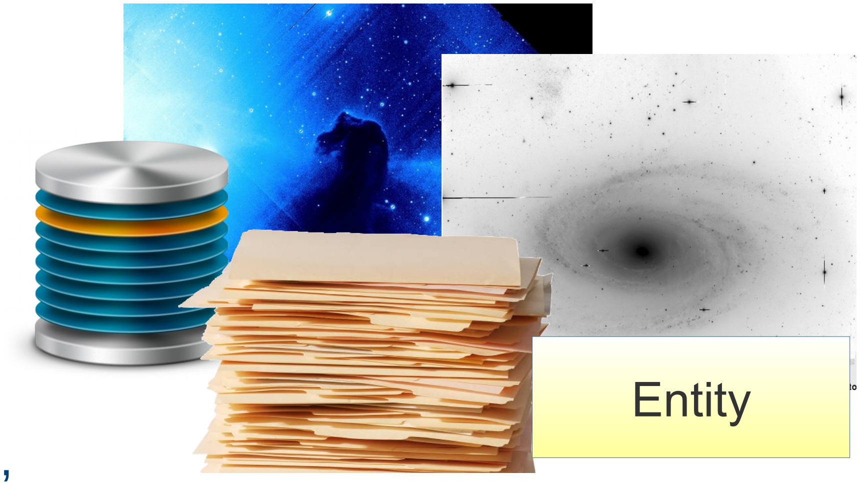


- identify data entities
- identify processes (activities)
- identify responsible people

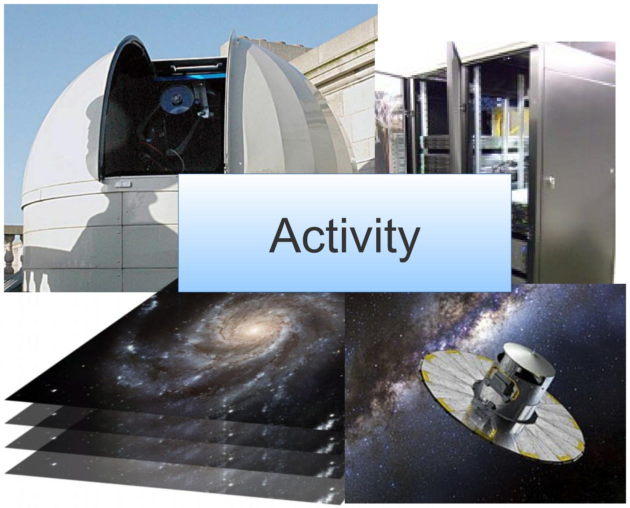
- provenance is defined by the relations between data, activities and the people/projects involved

Examples for core objects

- **Entities (datasets):**
  images, catalogs,
  database tables, spectra,
  log files, parameters, ...



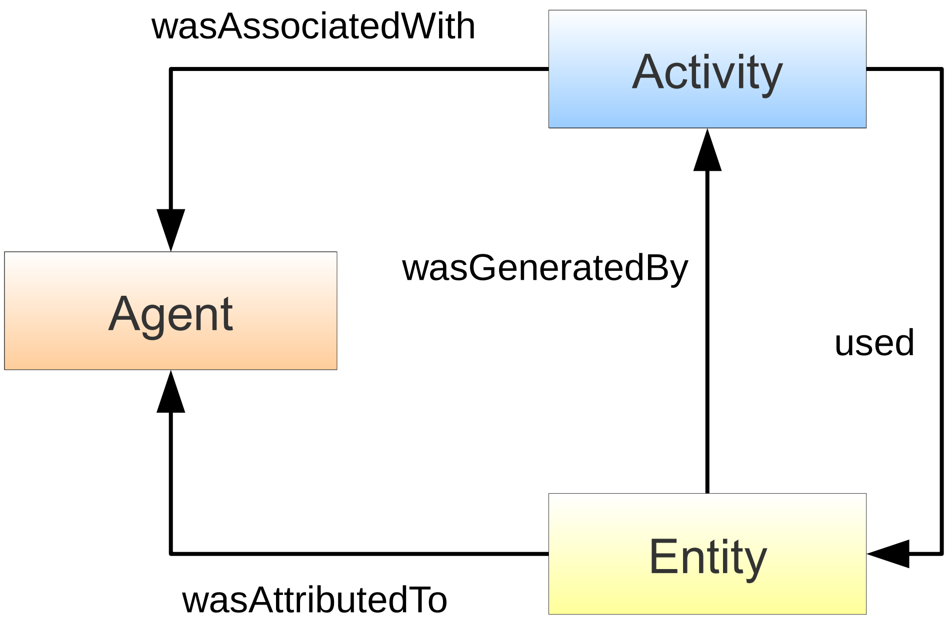
- **Activities:**
  observations;
  processing steps like bias subtraction, image stacking, continuum fit, object extraction; simulations, ...



- **Persons/Organizations:**
  creator, publisher, developer, ...

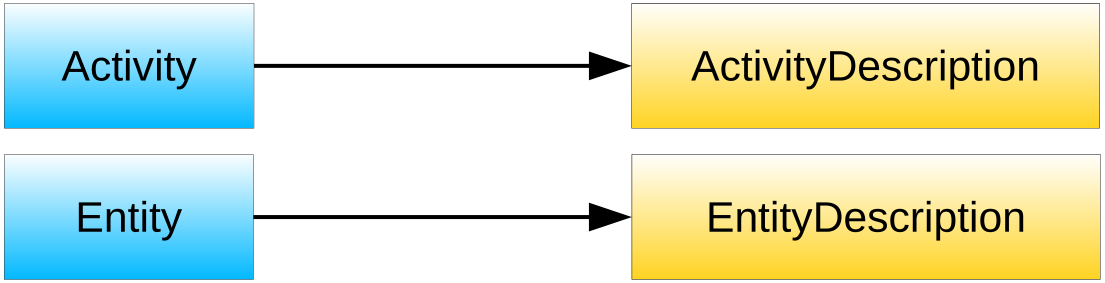
Provenance DM core classes

- 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



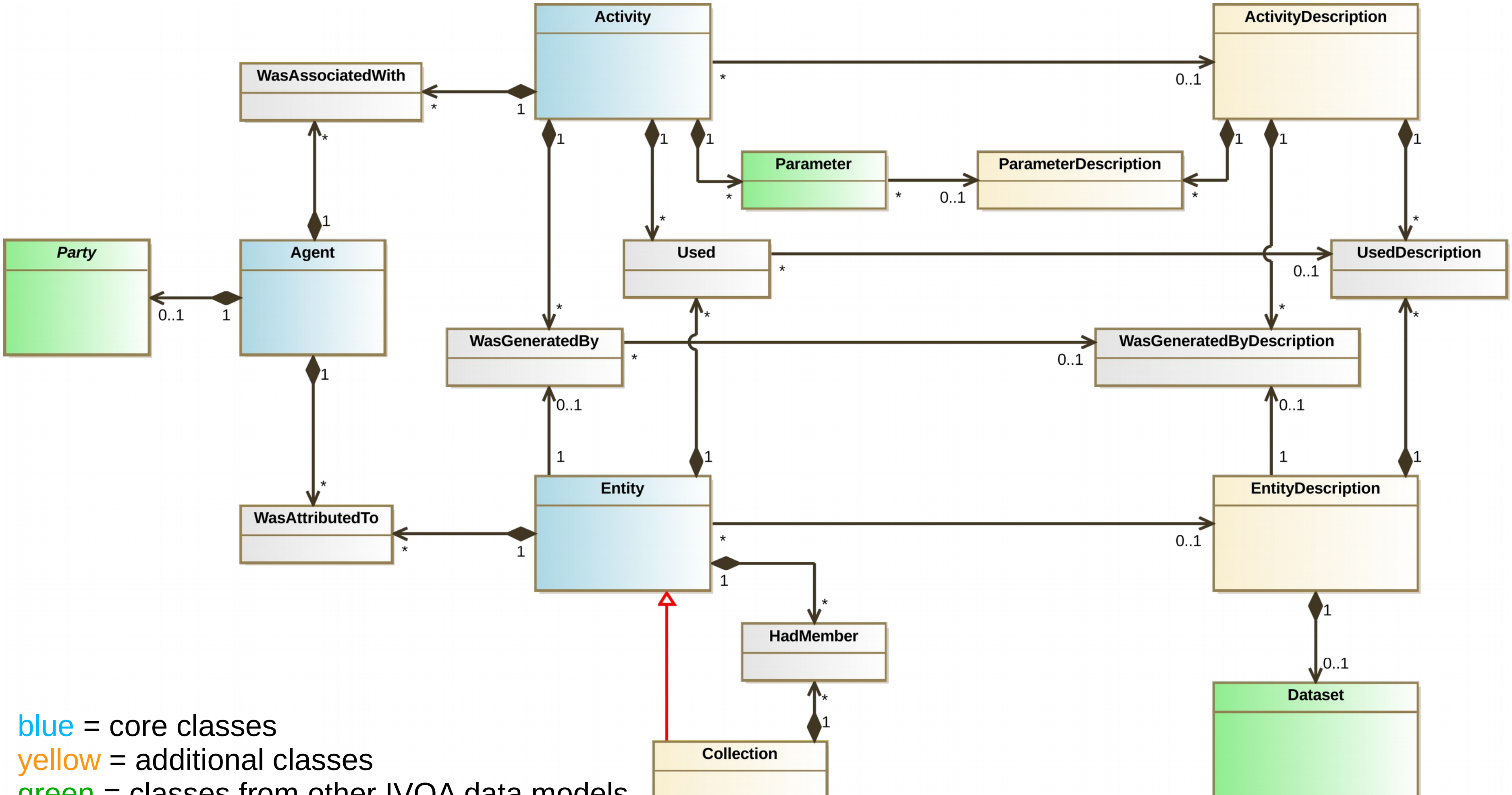
Extending the core

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



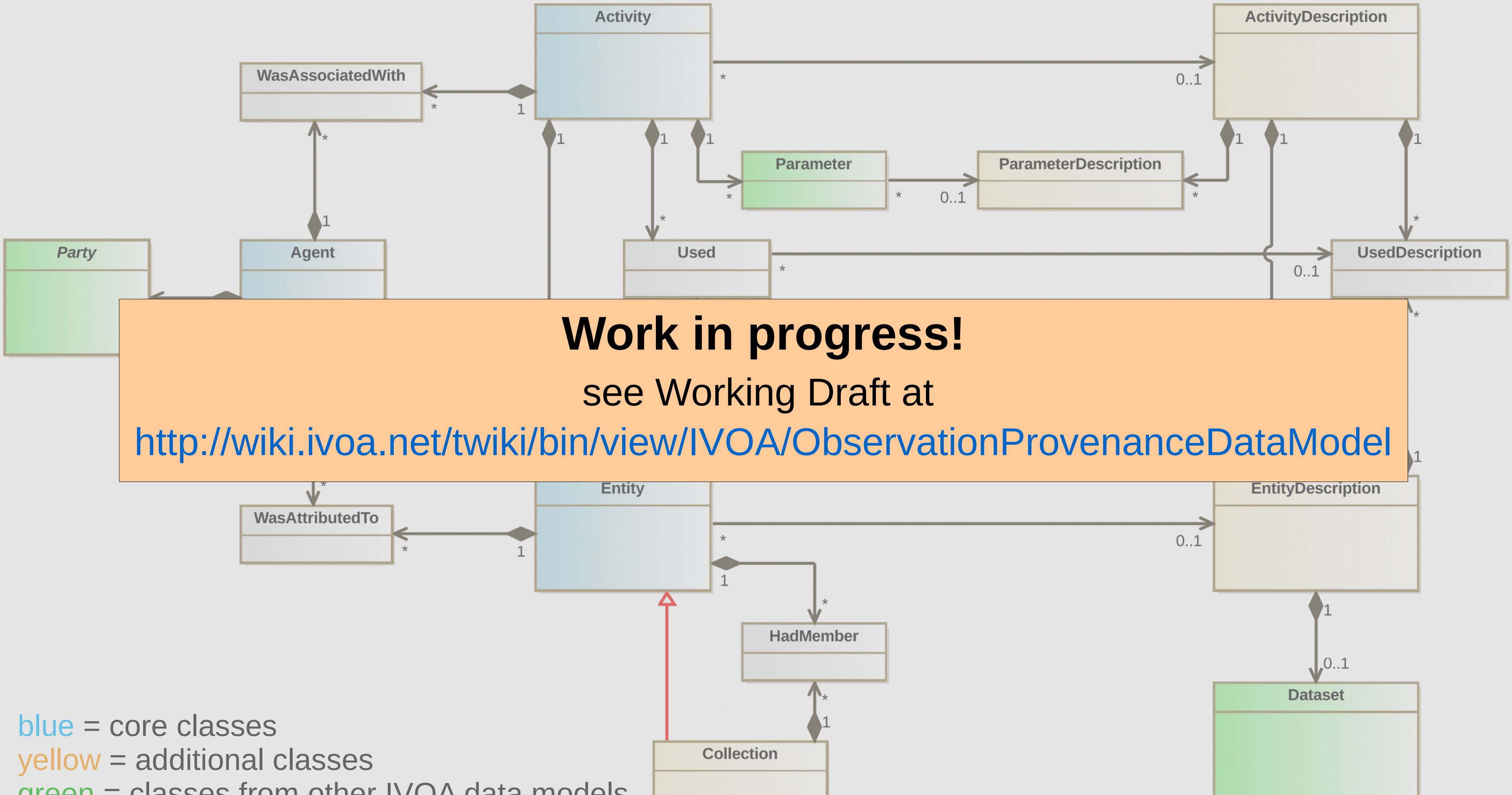
- connection to similar structures in other data models:
  - Activity => **Experiment** in Simulation Data Model
  - ActivityDescription => **Protocol** in Simulation Data Model
  - EntityDescription => **Dataset** in Dataset Data Model

Overview class diagram from working draft



blue = core classes
yellow = additional classes
green = classes from other IVOA data models
grey = relation classes

Overview class diagram from working draft



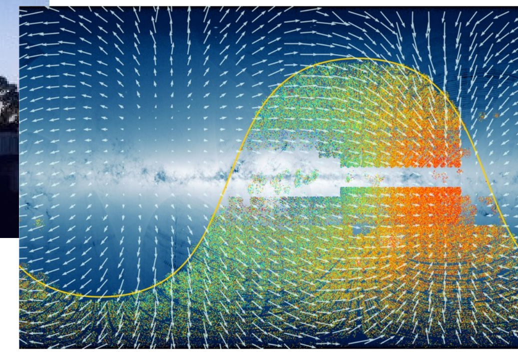
Work in progress!

see Working Draft at

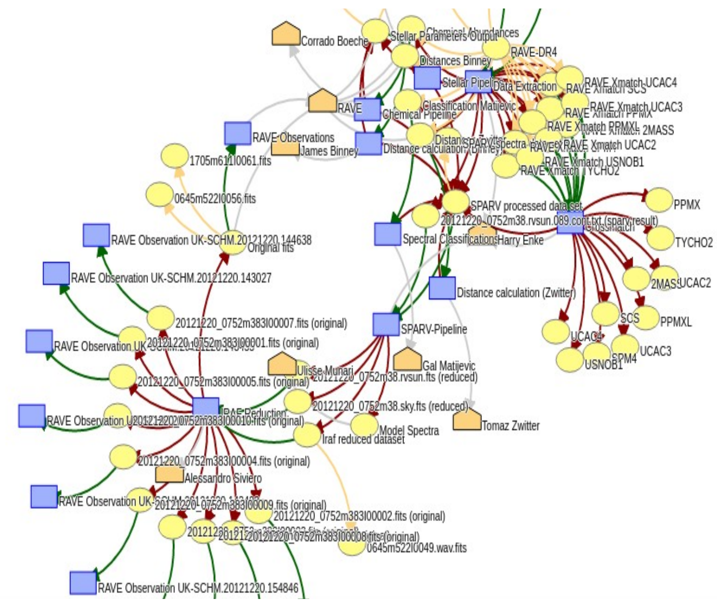
http://wiki.ivoa.net/twiki/bin/view/IVOA/ObservationProvenanceDataModel

blue = core classes
yellow = additional classes
green = classes from other IVOA data models
grey = relation classes

Use case: RAVE

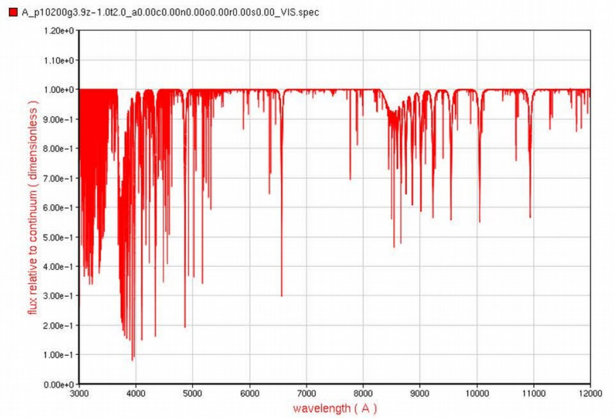


- multi-fibre spectroscopic survey
- radial velocities + derived stellar properties for ~ half million stars
- use provenance to track e.g.
  - Who was responsible for determining the log g values in DR5?
  - Which fibre observed the spectrum for star xyz?
  - Study selection effects using information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

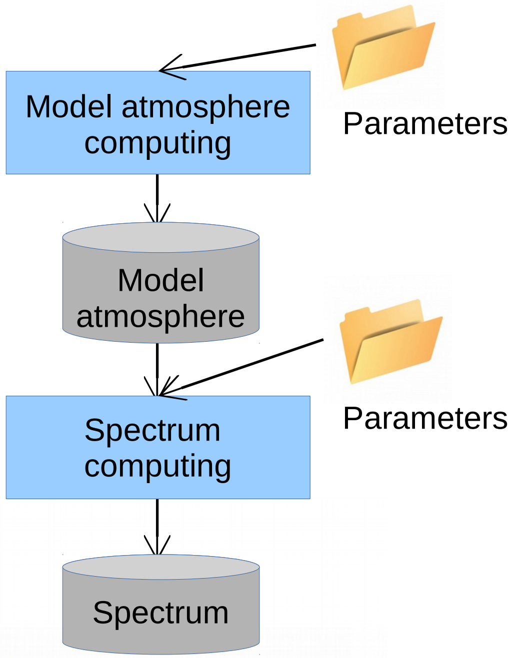


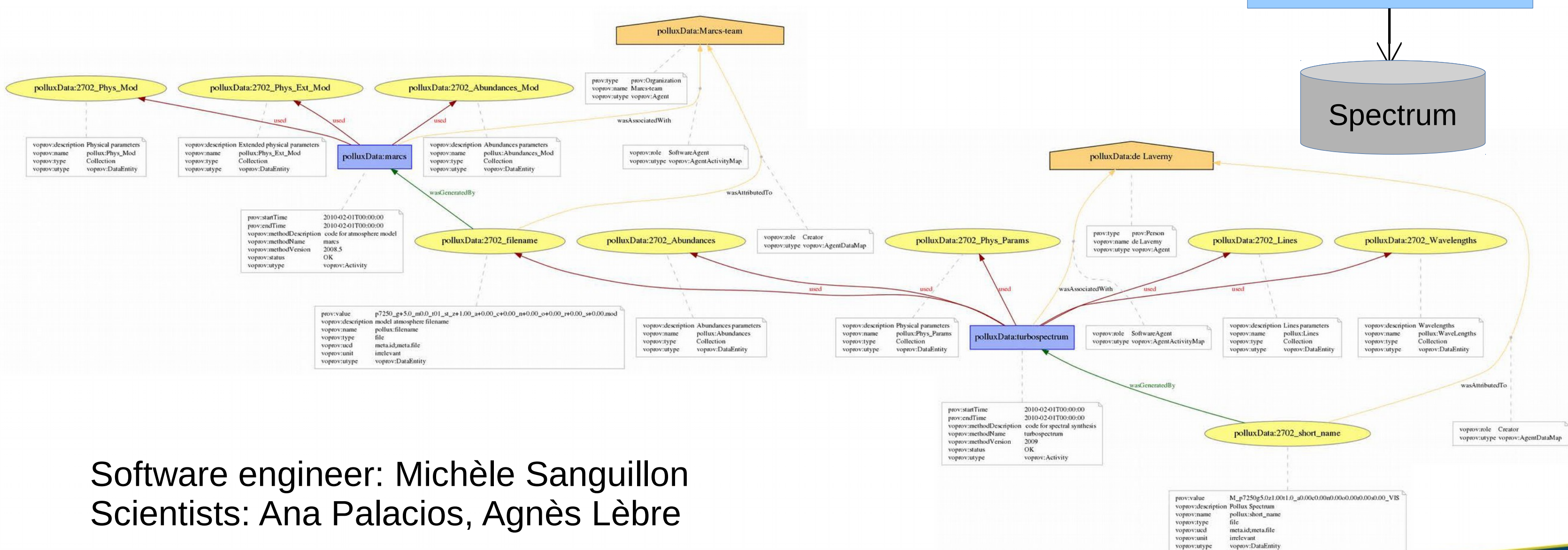
see javacript example at
https://escience.aip.de/prov/graphs/example.html

Use case: Pollux



Database of more than 8000 very high resolution synthetic spectra in optical domain
(3000 Å- 12000 Å)



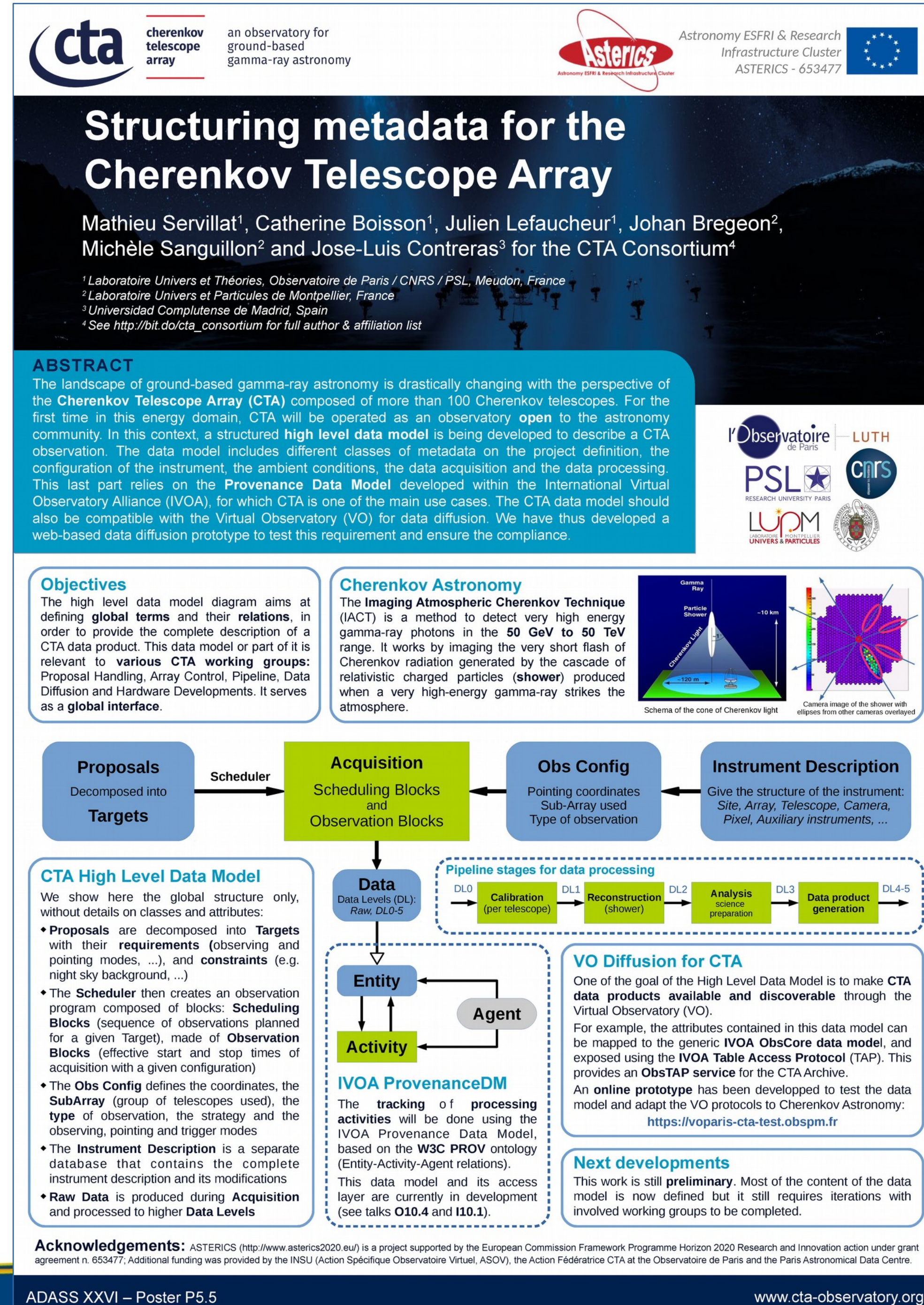


Software engineer: Michèle Sanguillon
Scientists: Ana Palacios, Agnès Lèbre

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 proccessing steps



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!