Provenance Data Model

Thoughts from GAVO

Kristin Riebe, Leibniz-Institut für Astrophysik Potsdam

Motivation

- What is provenance for?
- For a given data set, it should help to …
  - discover steps of production
    Which processing steps have been done already?
  - give attribution
    Who was involved in the project? Who can I ask about these data?
  - aid in reprocessing
    But not necessarily: allow reprocessing on keypress
  - aid in debugging
    Find possible error sources, e.g. check version of processing software, ambient conditions, telescope configuration, parameter settings, …
  - allow to assess the "quality" of the observation/processing
    → Quality DM?
  - search in structured provenance metadata

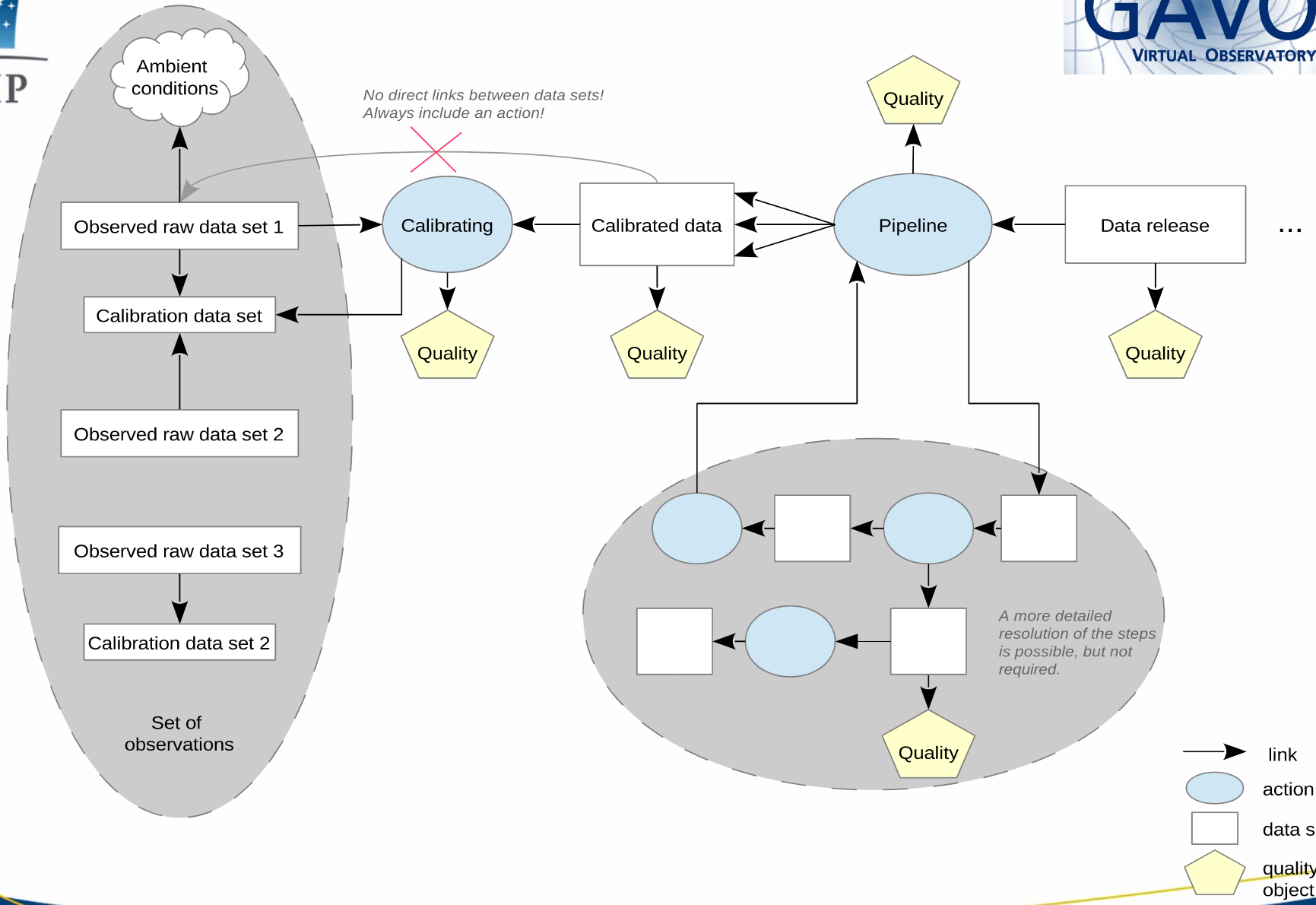
So far …

- collected some raw use cases to define the scope (MuseWise, RAVE, CALIFA, STELLA, also see talk by François Bonnarel last year)
- list of requirements for data model
- started with a list of processing steps that should be covered
- research on ambient conditions and instrumental parameters: trying to find common keywords

Requirements

- distinguish two types of things:
  data sets and **actions** (processes)
- links between data sets involve an action in between
- actions have input/output data set; provide links to them
- provide 'backward' links, from result to previous action/data set

Provenance - Example



Requirements II

- provenance data model should also cover
  - processes with/without raw data
  - non-automatic steps
    e.g. line fitting ‚by eye', awk/sed replacements in the bash, masking of foreground stars, …
- include ambient conditions, telescope, telescope site

  (data characterization => Characterisation Data Model; but things above not included ...)
- include observer/data creator + affiliation for reference

Processing steps I

- Data reduction
  - for CCDs: bias subtraction, dark field and flat field, rebinning of pixels
  - sky subtraction
  - remove hot & bad pixels
  - stacking images (reduce signal-to-noise ratio)
  - cosmic ray rejection
  - correction for atmospheric extinction, galactic extinction
  - spectra: flux calibration, wavelength calibration, correction for differential atmospheric refraction (DAR), image reconstruction
  - astrometric calibration

Processing steps II

- Data analysis
  - masking (e.g. of foreground stars or masking quasars to find quasar host galaxies)
  - for stars: fit point-spread function, convolve with it
  - combining signals (interferometry, radio telescopes like LOFAR)
  - cross matching with other catalogues
  - source extraction (e.g. with SExtractor, find stars, extended sources etc.)
  - spectra:
    - correct for redshift (from characteristic lines)
    - fit continuum
    - fit model atmospheres
    - fit synthetic spectra (to determine stellar parameters)

Ambient conditions

- What ambient conditions should usefully be covered by the PDM?
- Leech work done by designers of existing FITS headers: All-VO searches for Spectra and Images, extract headers
- (Interested? Want to contribute? Ask us!)
- Concept groups we've identified:
  - Geometry of celestial objects (e.g., SUNANGLE, DAYNIGHT, MOONFRAC, SUN_ALT...)
  - Atmosphere (AIRMASS, ZD)
  - Near-Instrument environment (e.g., temperature, pressure – demarcation to instrument telemetry not always clear)
  - Environmental Hazards (e.g., „LWR header warmup“ – demarcation to instrument telemetry and process description not always clear)
  - Sensor location and movement (e.g., SITELAT, SITELONG, ORBAXIS, V_GEOCEN, INCLINAT...)
  - „Freetext“ (QUALCOMi, QUALITY)

FITS keywords: Lessons learnt

- All told, we've collected about 50 FITS keywords we'd put into the ambient condition group.
- For instrument metadata, our small sample already has about 700 FITS keywords.
- Clearly, these cannot be directly mapped into data model components (even if there were a use case for them)
- Proposal:
  - Simple DM
    (e.g., conceptName, conceptValue, valueUnit, relation*)
  - have concepts in a thesaurus, including wider/narrower relations, where terms never die.

Questions I

- Allow to group processing steps?
  - How?
  - Benefit: different layers of „resolution“; if storing provenance information in fits-header, it can be easier to handle coarser information, which could be looked up in detail at a „provenance repository“
- Workflow management systems (e.g. AstroTaverna):
  - Could use their experience, what did they include? What is missing?
  - easily track workflow and thus provenance of a data set
  - => follow each step? (or at least link to AstroTaverna‘s log?)
- Access
  - allow restricted access?

Questions II

- How to treat „political“ information?
  - e.g. project name, PI of project, link to proposals
  - partially given in fits-headers
  - could be used for linking telescopes with scientific outcome/impact

    => Should it be included in Provenance Data Model or is it out of scope?

- Implementation
  - How and where to store this information? Provenance repository similar to VO registry?
  - Keep as much information with the data as possible?