Prototyping Provenance metadata for the Virtual Observatory

M. Louys, F. Bonnarel, A. Nebot, CDS

C. Boisson, M. Servillat, M. Sanguillon, J. Bregeon, CTA

L. Michel, High Energy Group, Strasbourg Observatory

Provenance metadata 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, …)
- 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.

  →Expose progenitors of science data products

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 assessments about its quality, reliability or trustworthiness. PROV-DM is the conceptual data model that forms a basis for the W3C provenance (PROV) family of specifications."

PROV-OVERVIEW (Note), an overview of the PROV family of documents

PROV-PRIMER (Note), a primer for the PROV data model

PROV-O (Recommendation), the PROV ontology, an OWL2 ontology allowing the mapping of the PROV data model to RDF

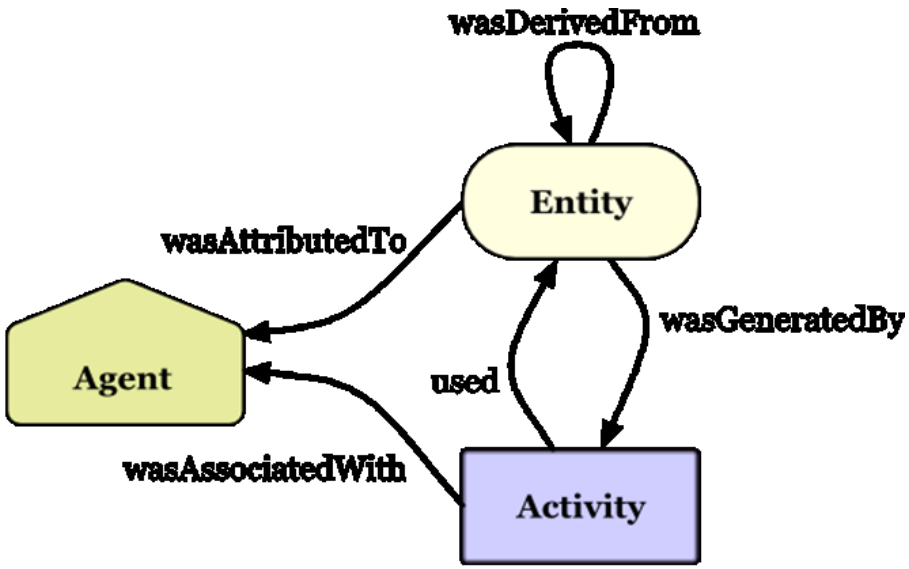
PROV-DM (Recommendation), the PROV data model for provenance (this document)

PROV-N (Recommendation), a notation for provenance aimed at human consumption

PROV-XML (Note), an XML schema for the PROV data model

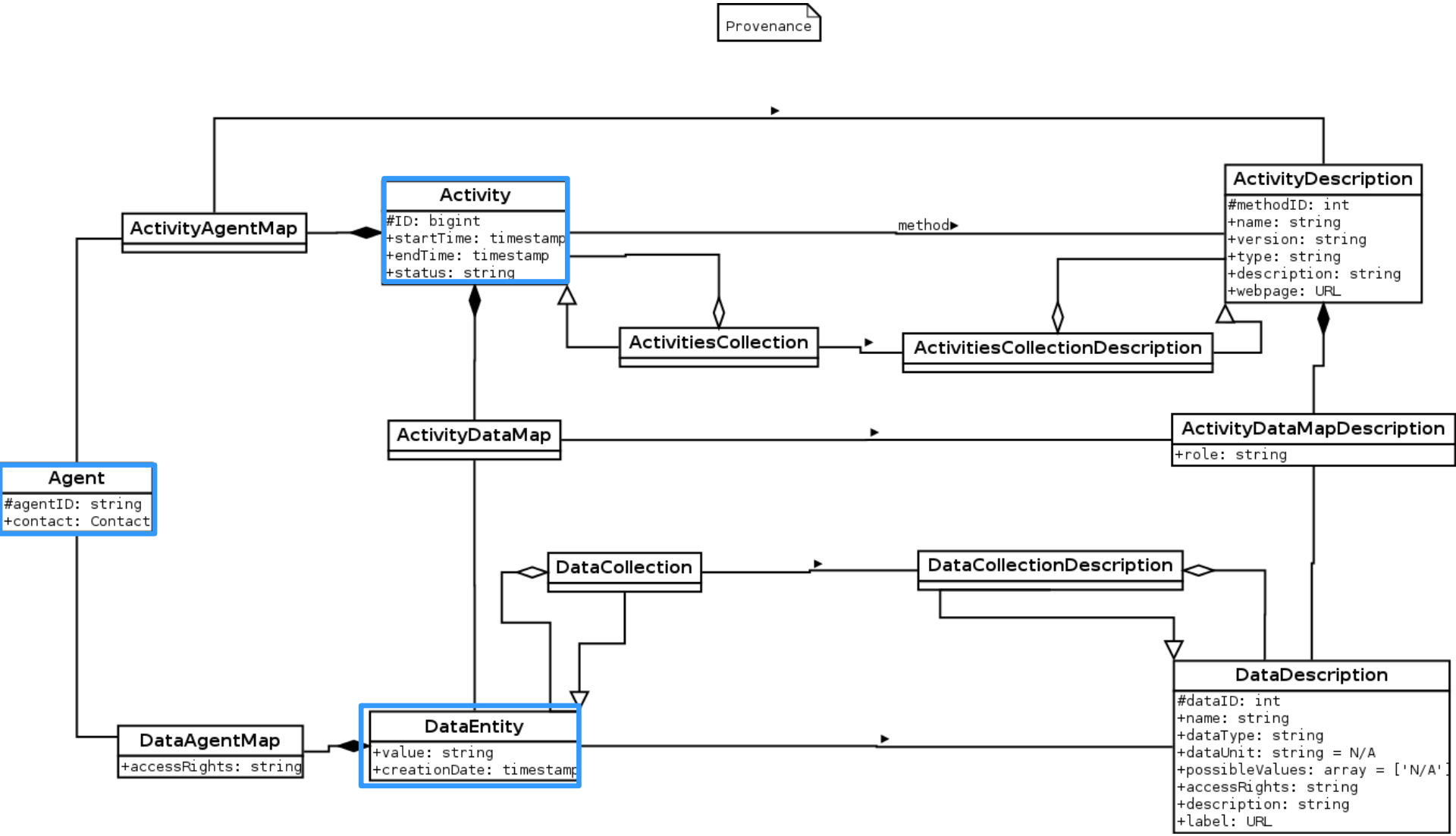
PROV-AQ (Note), mechanisms for accessing and querying provenance

W3C Provenance pattern



- Makes explicit:
  - Processing steps
  - Chain of dependencies
  - Responsibilities
- Useful for all execution sequence of tasks, workflows, reduction pipeline, analysis workflow, etc.
- Applies for both acquisition and reduction steps

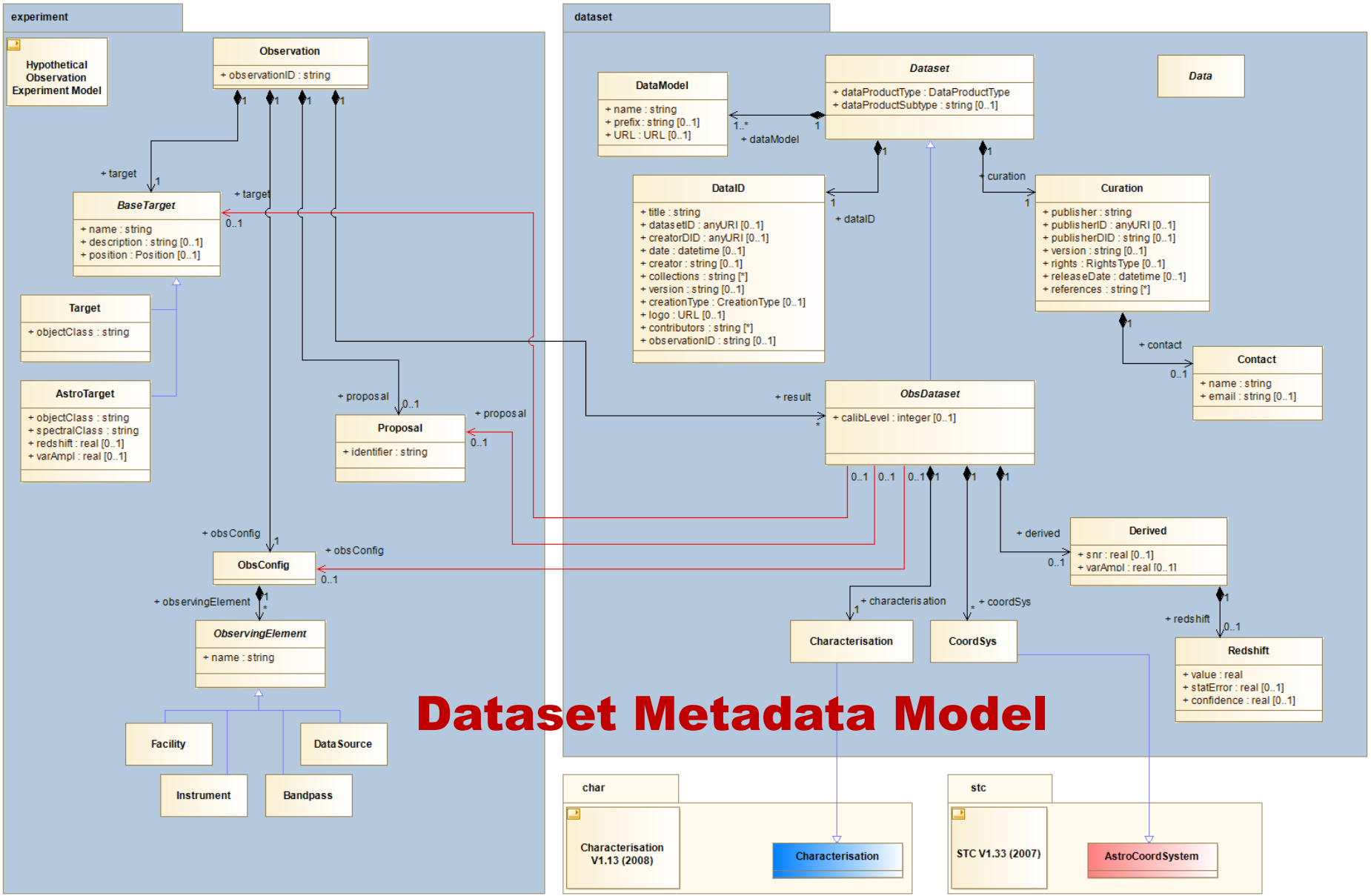
Customized re-use



In our context

- Entity = data products (files), ancillary data (calibration, psf, instrumental response, etc), parameter files
- Activity = data acquisition, regridding, fusion, calibration step , DL3 to DL4 (CTA)
- Agent = telescope astronomer, pipeline operator, principal investigator, etc.

Binding to existing data models



More to do

- More use-cases to work out for the data models combination
- Explore the ActivityDescription class for various use-cases
  - M.Servillat, C. Boisson, M.Sanguillon , J. Bregeon

    → CTA data products (4 levels of progenitors)
  - A. Nebot , L. Michel

    → Fitting parametric models profiles for XMM spectra
  - M. Sanguillon, LUPM , Montpellier

    →Provenance of theoretical spectra for the Pollux data base

Need for a serialisation format

- Currently most of the provenance information is available as :
  - log files
  - list of launched command lines in FITS headers in COMMENT keywords
- W3C offers various forms of syntax, translators
- see Kristin Riebe Rave use-case http://wiki.ivoa.net/internal/IVOA/InterOpJune2015DM/Provenance.pdf
- PROV-N
  - Traces the execution scenario
  - Defines a grammar
  - Is simple text

Future plans

- Finalise prototypes for use-cases with
  - A light-weight model
  - Example scenarios with a list of real instances for ActivityDescription
  - Prov-N serialised example scenarios