DatasetDM: roadmap splinter

Discussion of Dataset model content in the current and future data model ecosystem.

Mark Cresitello-Dittmar

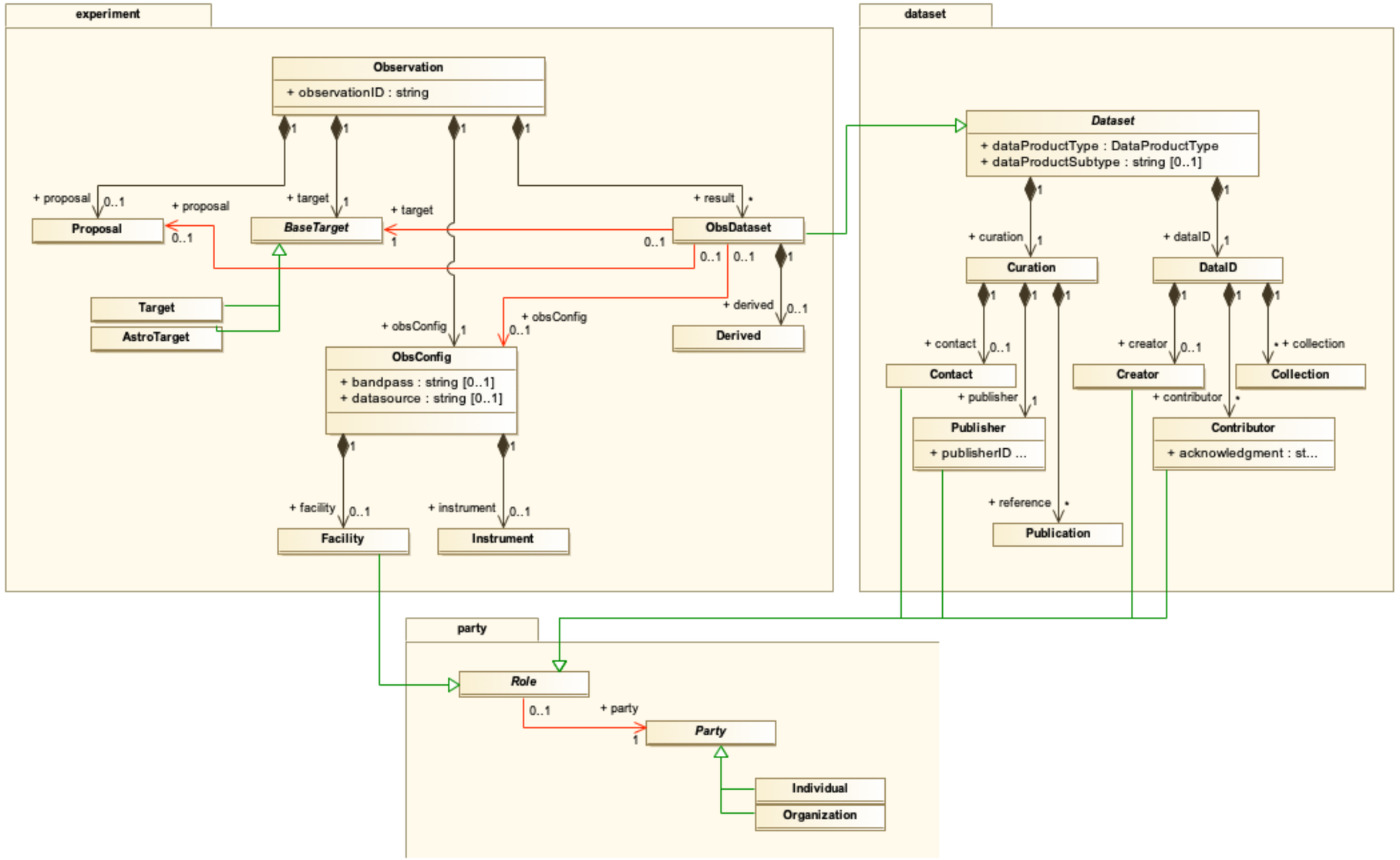
May 09, 2023

Topics of this discussion

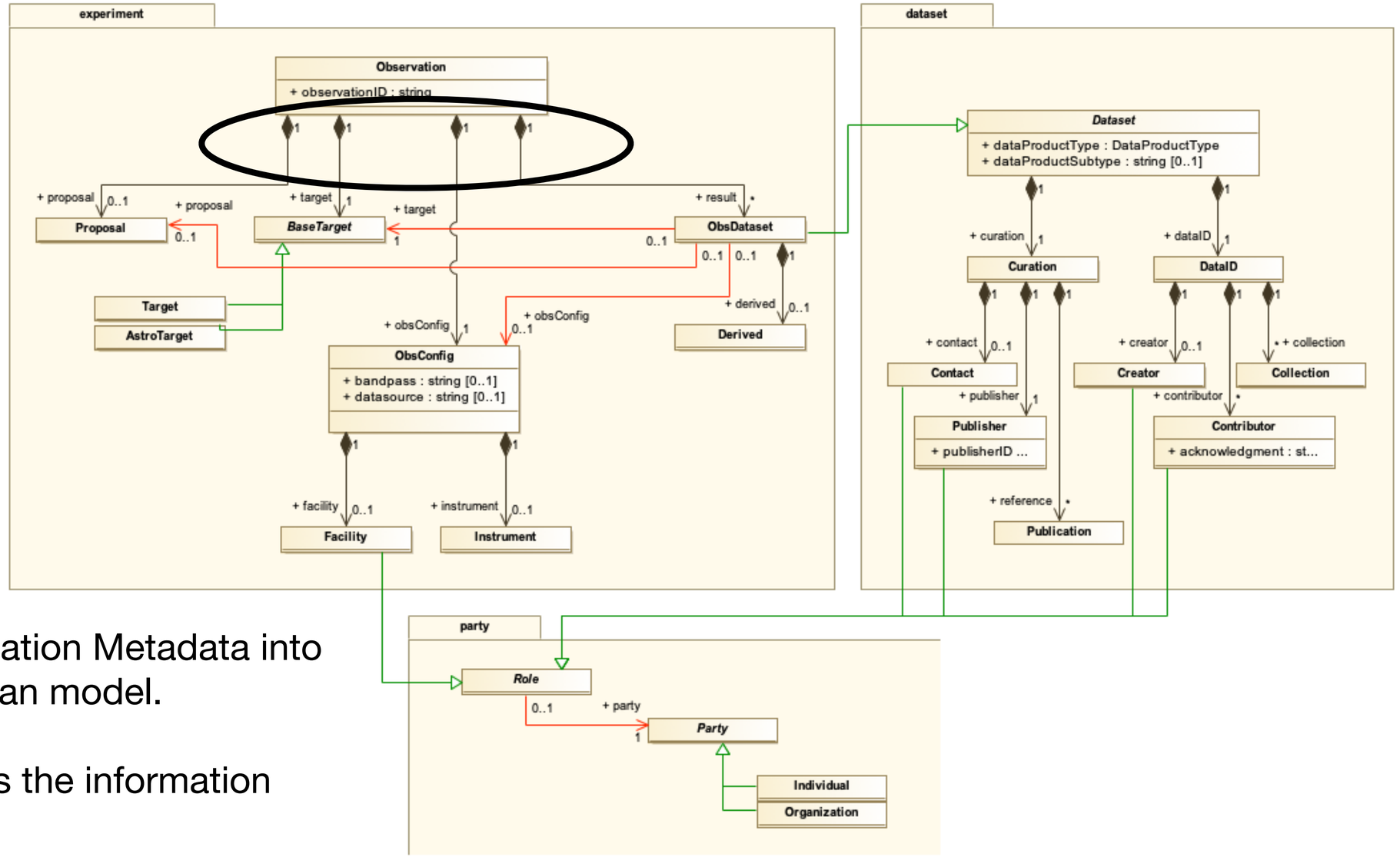
- Dataset Metadata model contains many concepts which essentially describe Provenance information pulled from the Activity tree leading to the generation of the Dataset itself (Entity).
  - I'd like to discuss how we should represent this information in the Dataset model moving forward.
- Many of these concepts are related to the Observation and these have been extracted to a straw man Observation model within the document. There is talk of migrating CAOM to an IVOA standard, since it is already being implemented widely within the community
  - It would be wise to make a plan, and work toward that while these overlapping models are being worked.
- Characterization model elements were part of the 'Dataset Metadata' content in the progenitor models (ObsCore, Spectrum). It was left behind, primarily because it is not VO-DML compliant.
  - It would be good plan the how/when this is folded back into the model.

Dataset DM

Overview



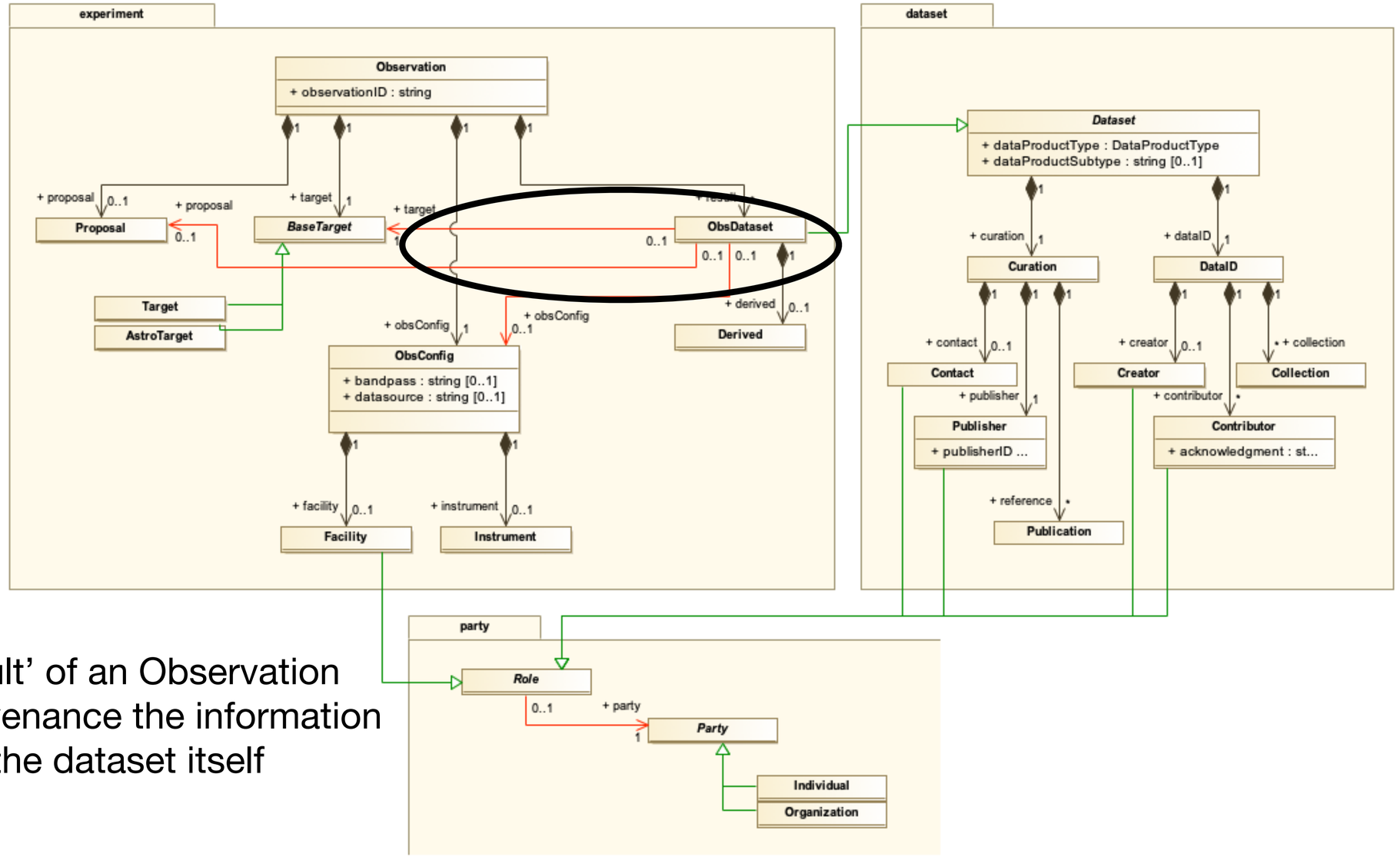
Dataset DM



Extracted Observation Metadata into separate straw man model.

Observation owns the information

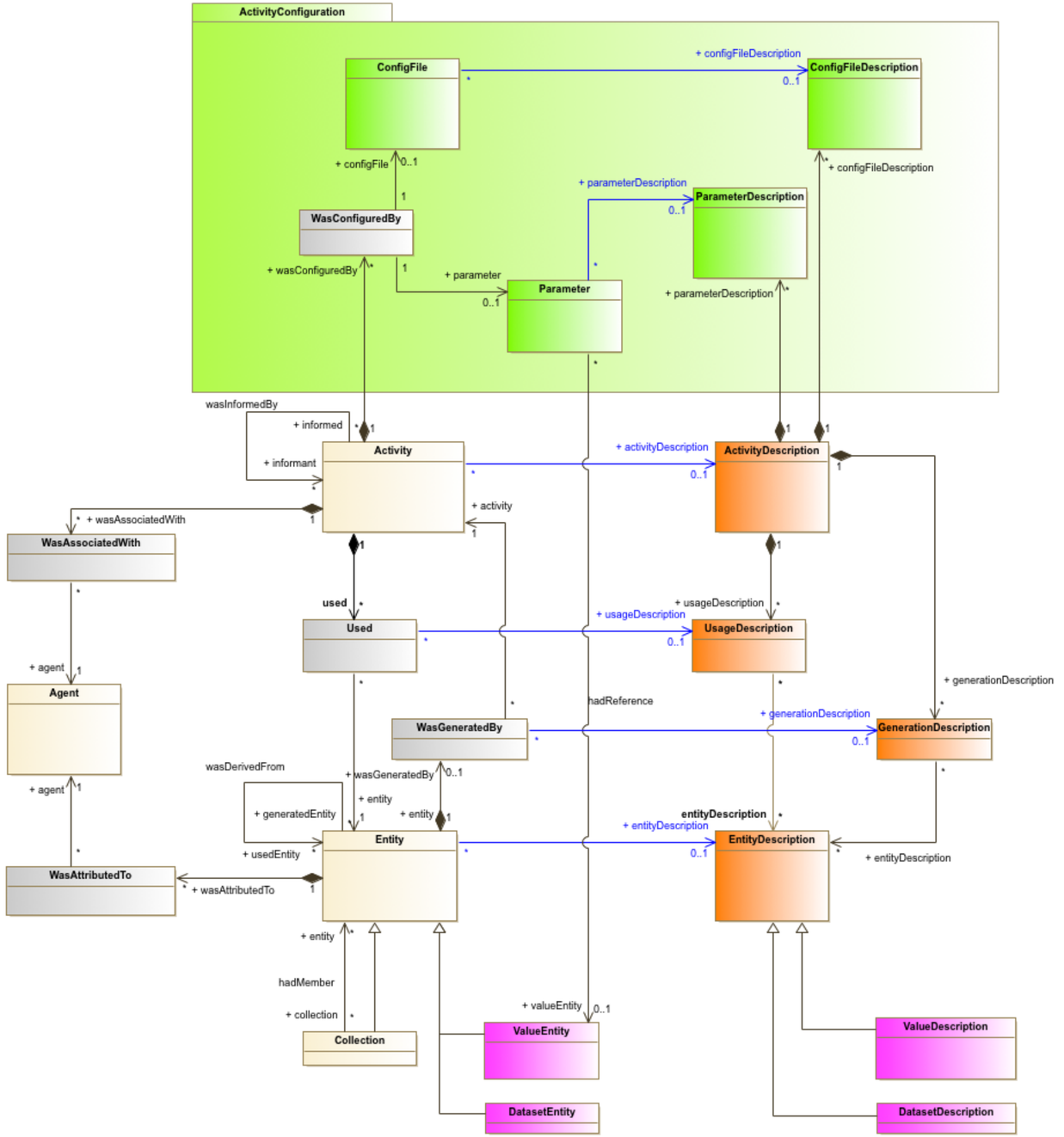
Dataset DM



ObsDataset is a ‘result’ of an Observation

* references as Provenance the information to be carried with the dataset itself

Provenance

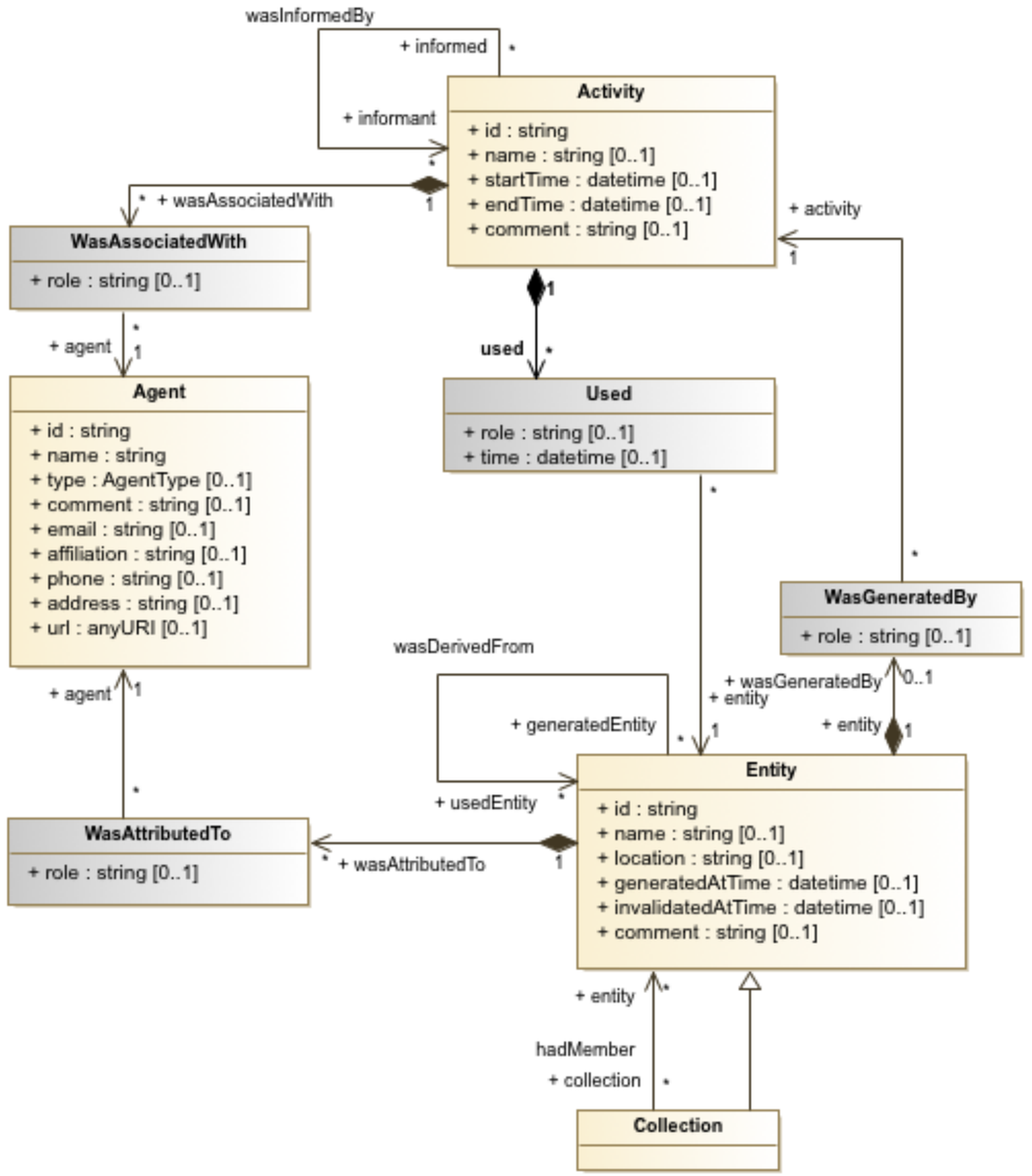

ActivityConfiguration
ConfigFile
+ configFileDescription
ConfigFileDescription
0..1
+ configFile
0..1
1
+ configFileDescription
+ parameterDescription
ParameterDescription
0..1
WasConfiguredBy
+ wasConfiguredBy
1
+ parameter
Parameter
0..1
+ parameterDescription
wasInformedBy
+ informed
+ informant
Activity
+ activityDescription
ActivityDescription
0..1
+ activity
+ wasAssociatedWith
WasAssociatedWith
used
Used
+ usageDescription
+ usageDescription
UsageDescription
0..1
+ agent
Agent
hadReference
+ generationDescription
+ generationDescription
WasGeneratedBy
GenerationDescription
0..1
wasDerivedFrom
+ wasGeneratedBy
0..1
+ entity
+ generatedEntity
+ entity
entityDescription
+ entityDescription
Entity
EntityDescription
0..1
+ usedEntity
+ agent
WasAttributedTo
+ wasAttributedTo
+ entityDescription
+ entity
hadMember
+ collection
+ valueEntity
0..1
ValueEntity
ValueDescription
Collection
DatasetEntity
DatasetDescription

Provenance

Agent - Activity - Entity

These same concepts are clearly evident in the dataset metadata.

How should we represent Provenance information in Datasets?

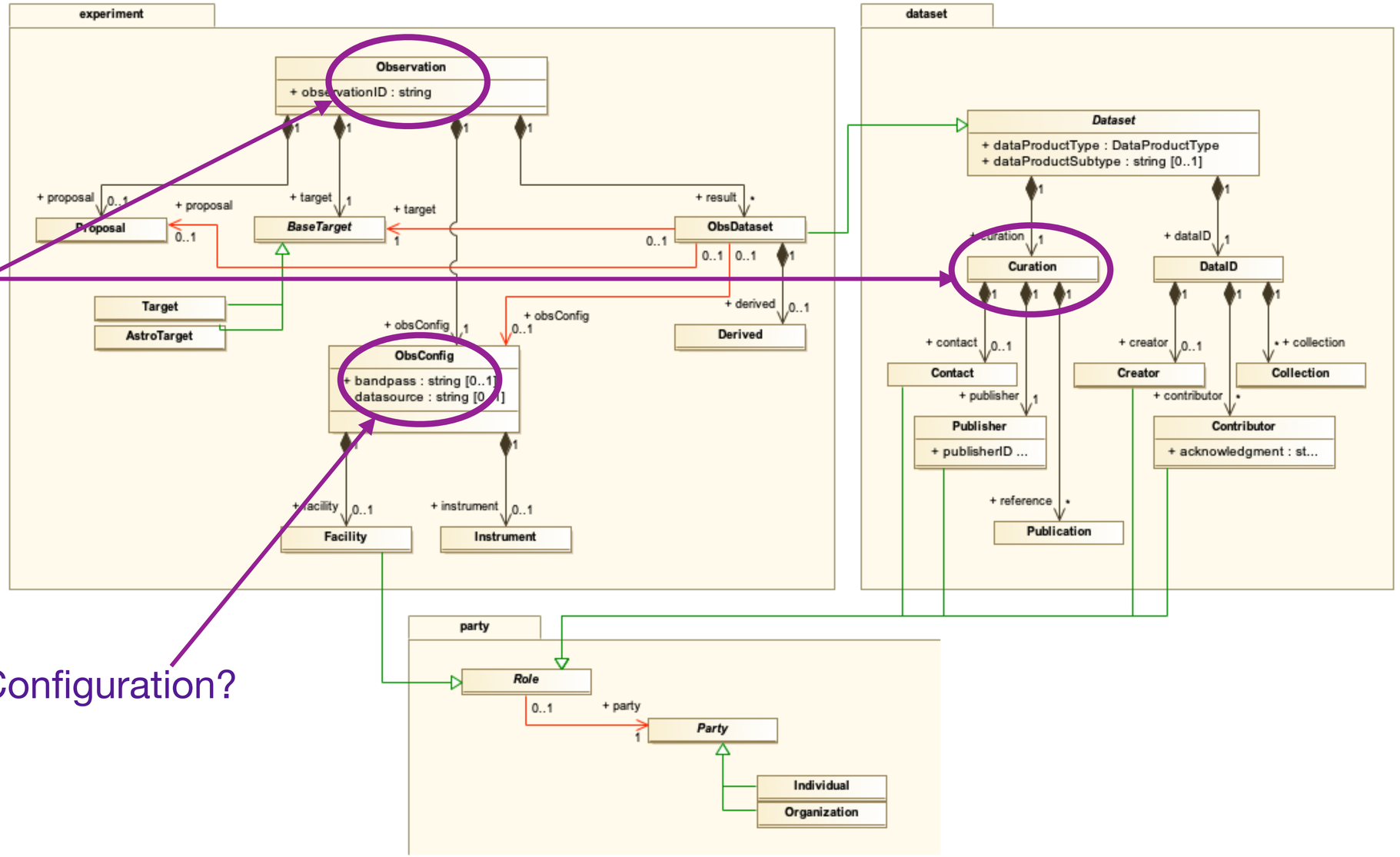


Dataset DM

Provenance

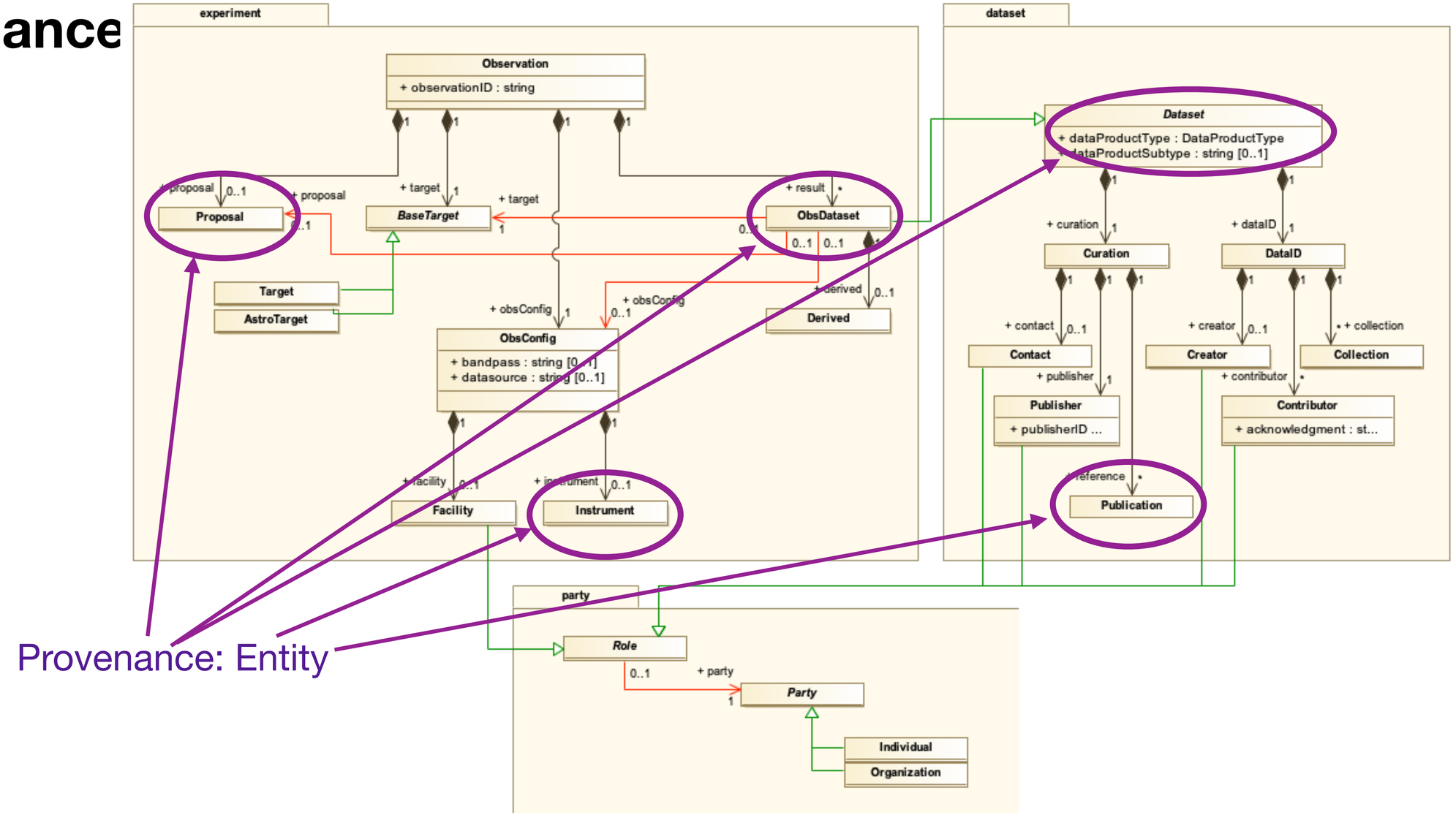
Provenance: Activity

Provenance: ActivityConfiguration?



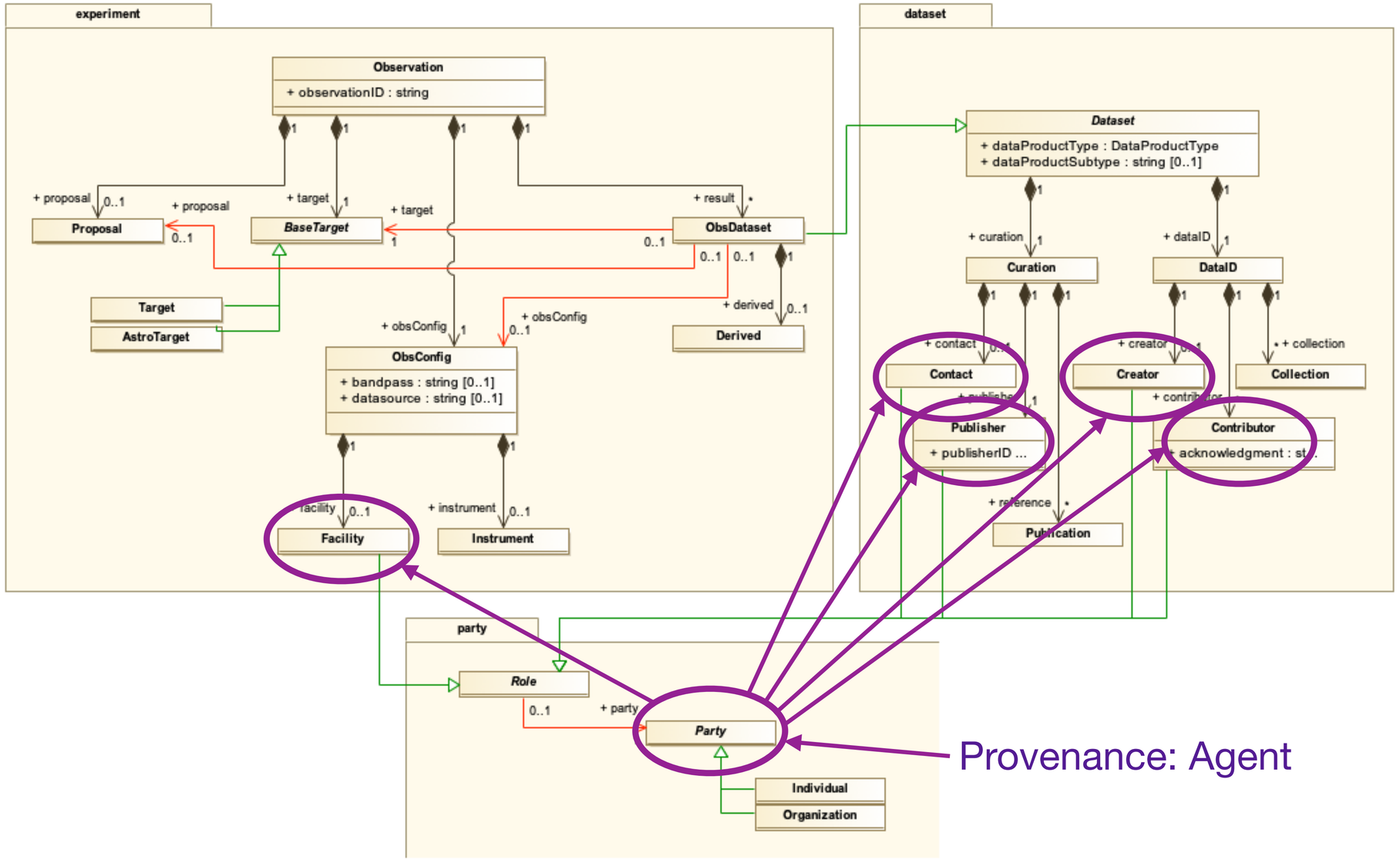
Dataset DM

Provenance



Dataset DM

Provenance



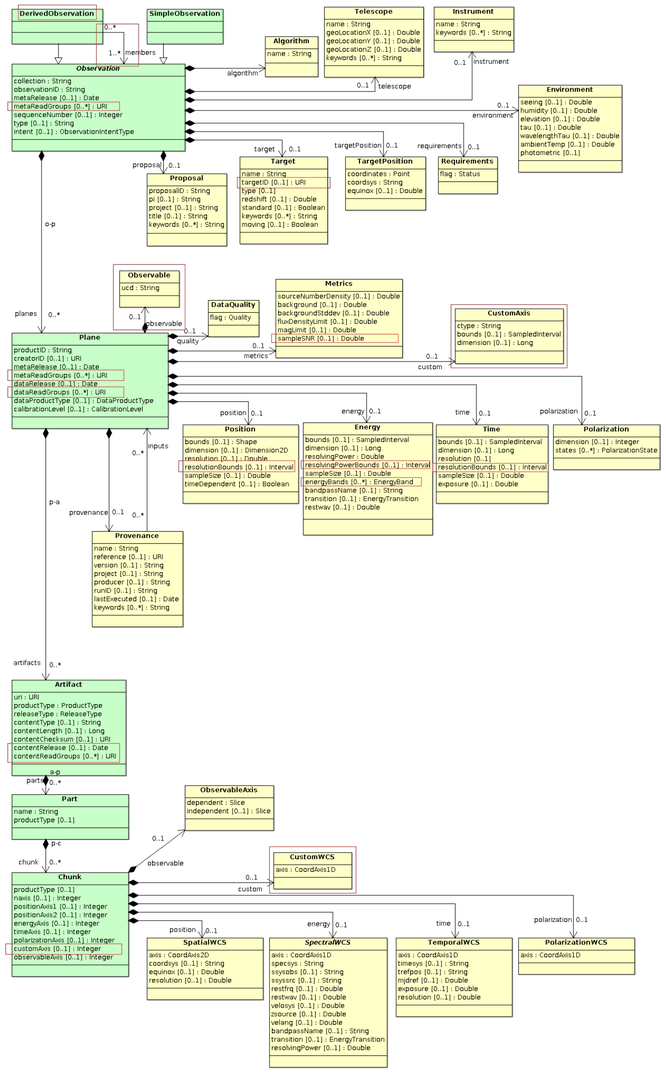
Provenance: Agent

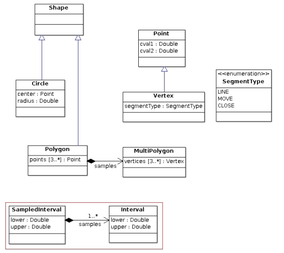
Dataset - Provenance relation

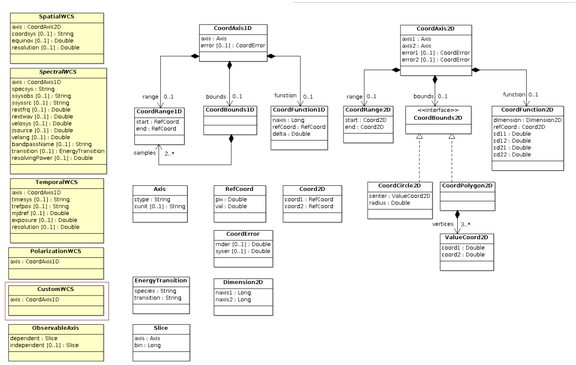
- How do we want to represent the various Agent - Activity - Entity elements in the Dataset DM?
  - What model pattern to use? Provenance involves references, which can be somewhat awkward in serializations. Can/should we use DataTypes?
  - Activity hierarchy? The Provenance pattern generates a hierarchy of Entities which is largely flattened in this model. For example, Created Dataset vs Curated Dataset.

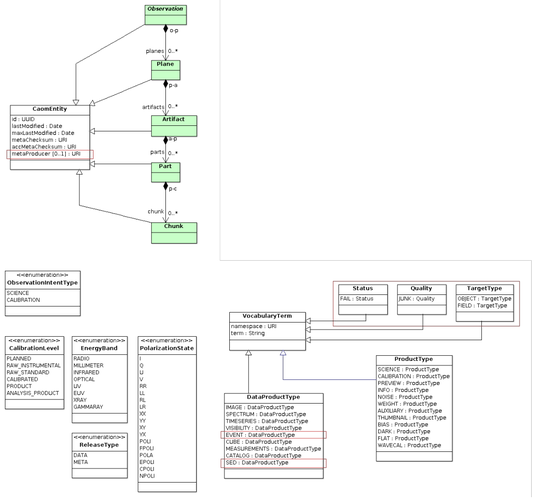
CAOM - Common Archive Observation Model

Overview





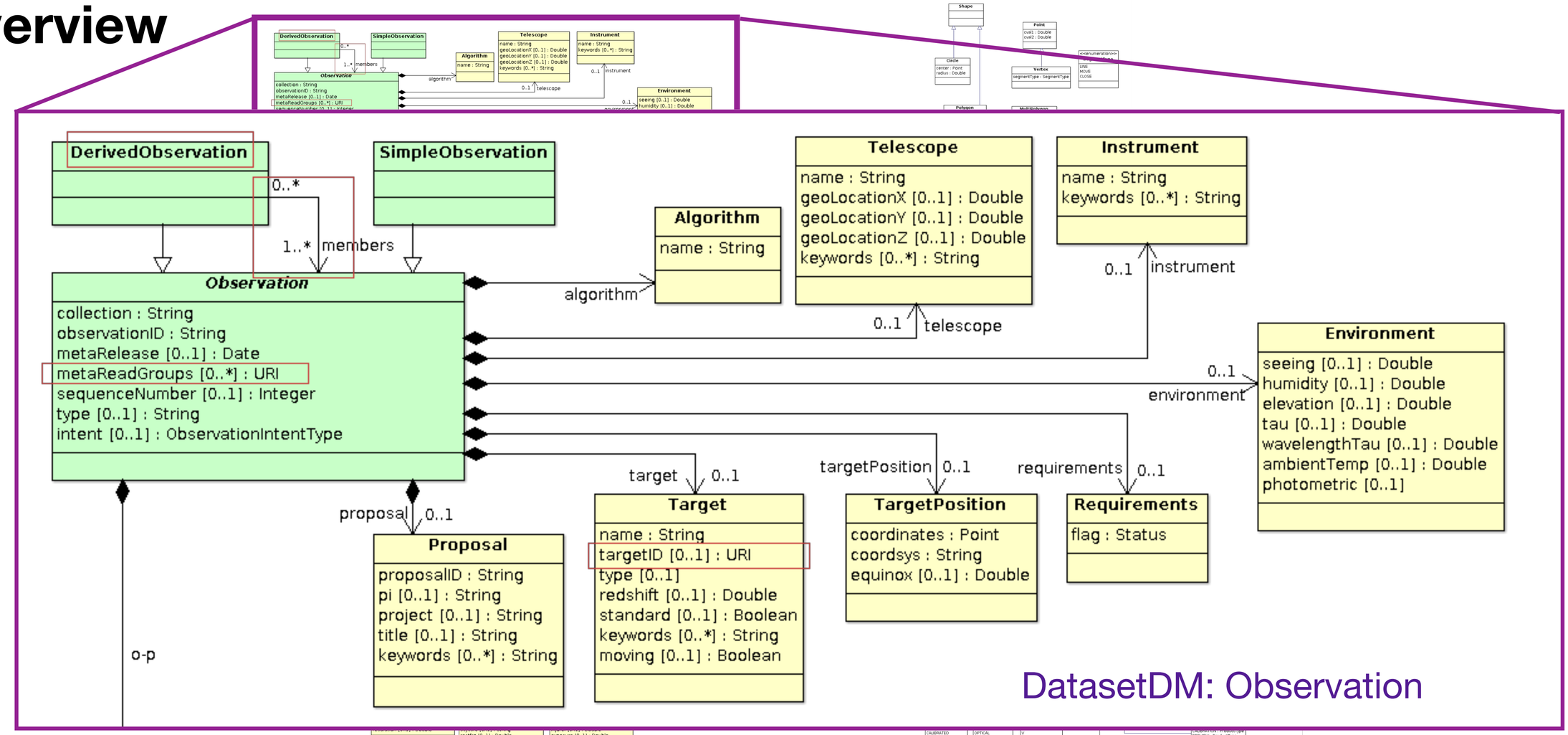




https://www.opencadc.org/caom2/

CAOM - Common Archive Observation Model

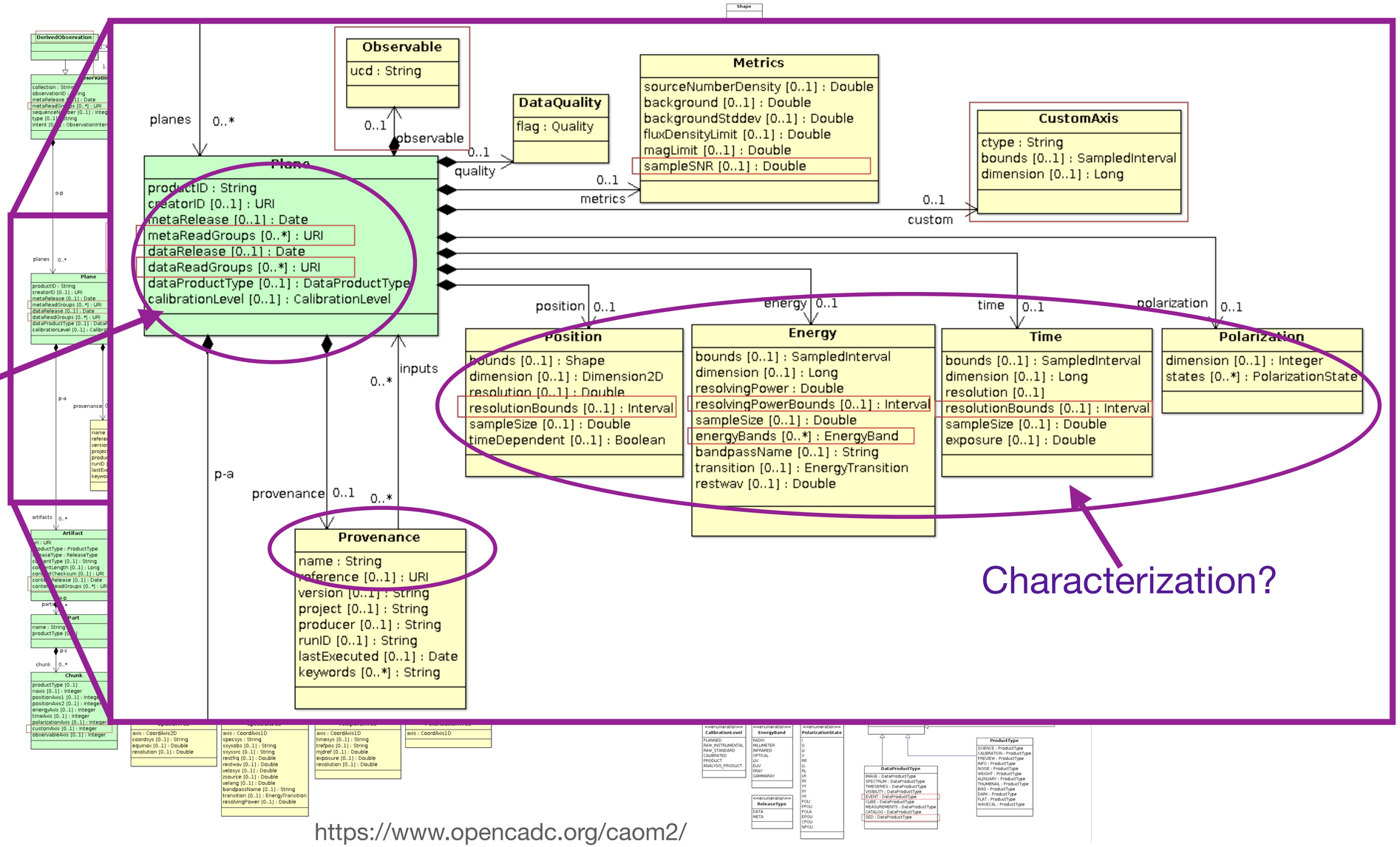
Overview



https://www.opencadc.org/caom2/

CAOM2 - Common Archive Observation Model

Overview



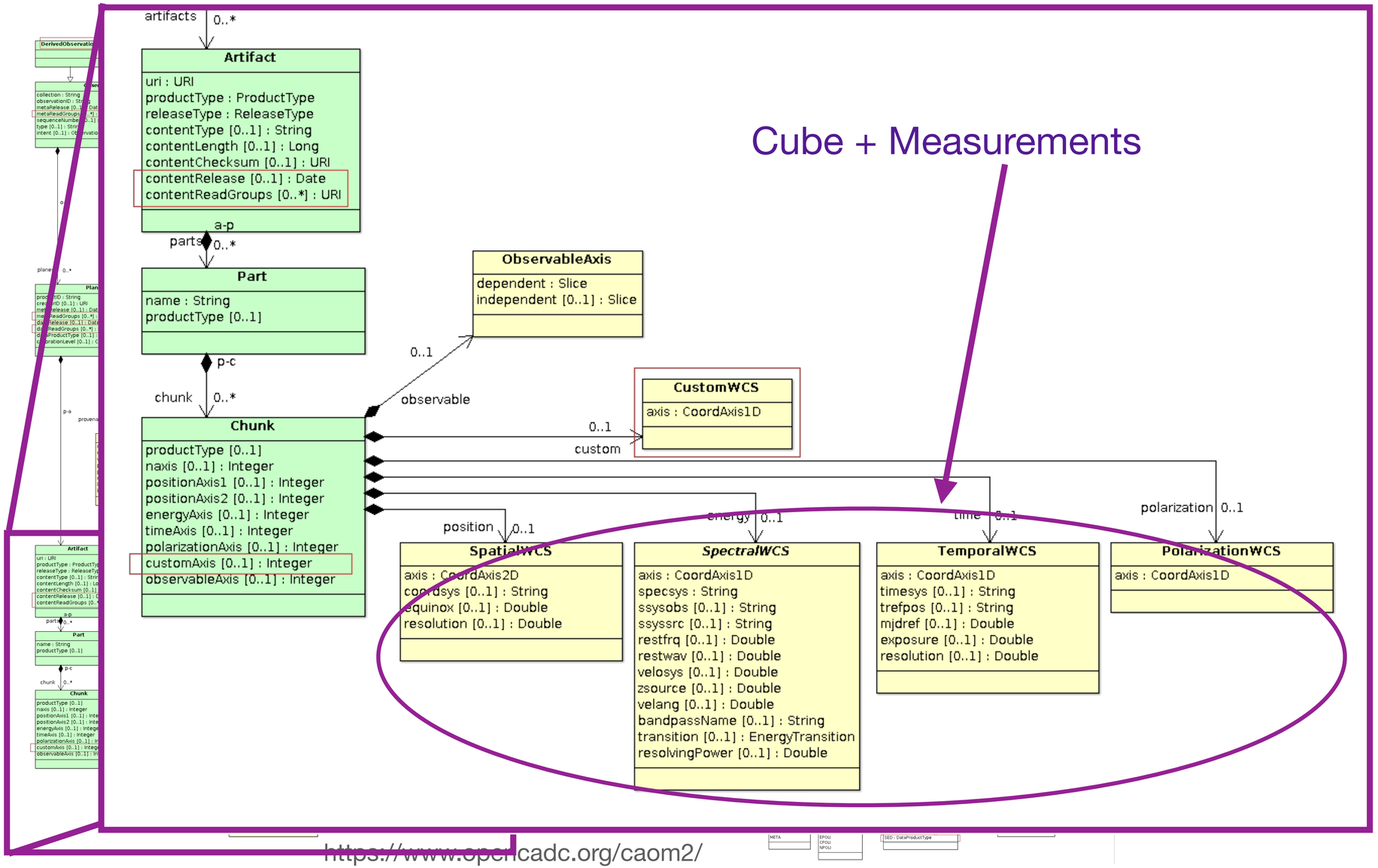
DatasetDM: Dataset

Characterization?

https://www.opencadc.org/caom2/

CAOM2 - Common Archive Observation Model

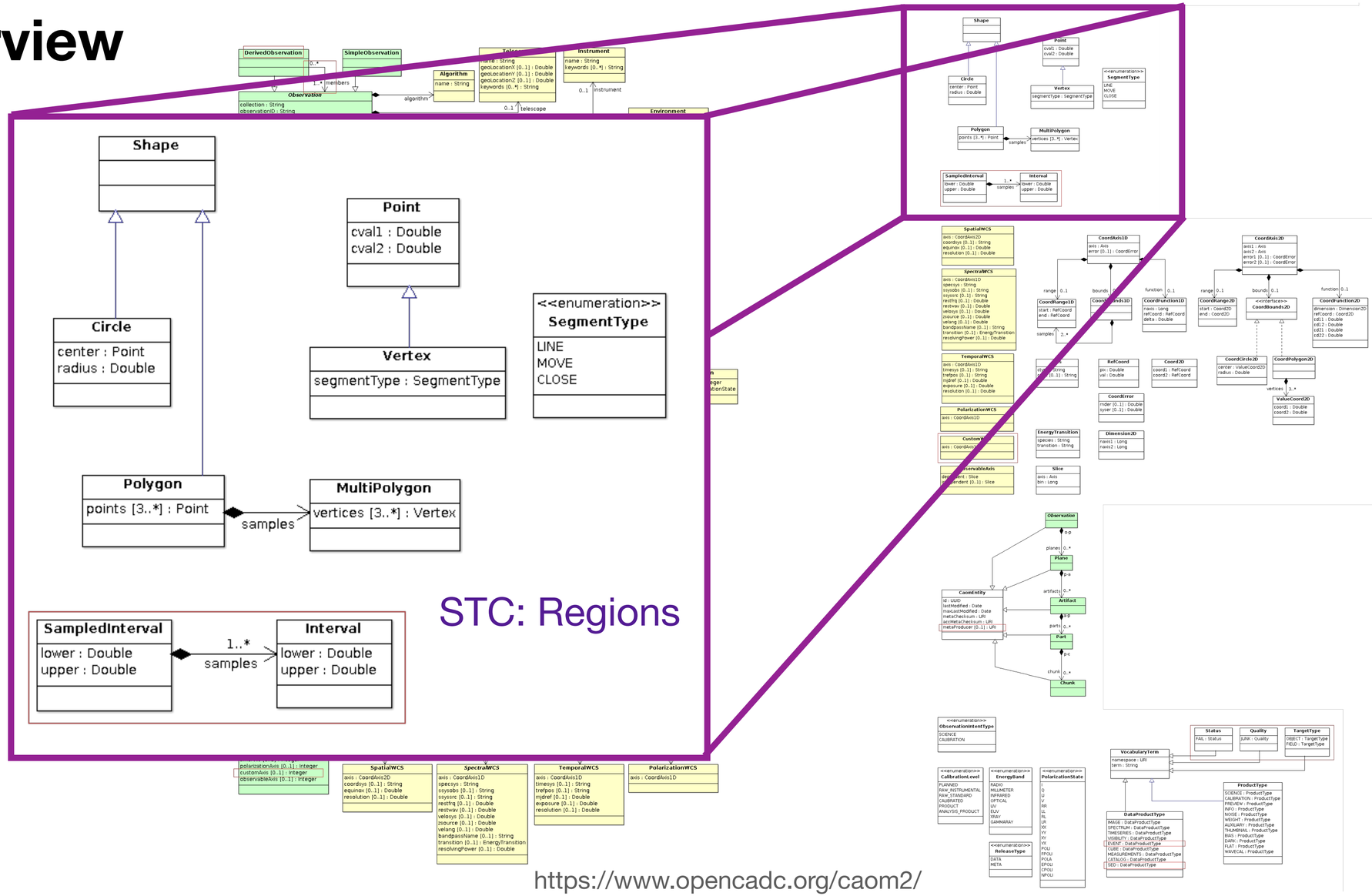
Overview



https://www.opencadc.org/caom2/

CAOM2 - Common Archive Observation Model

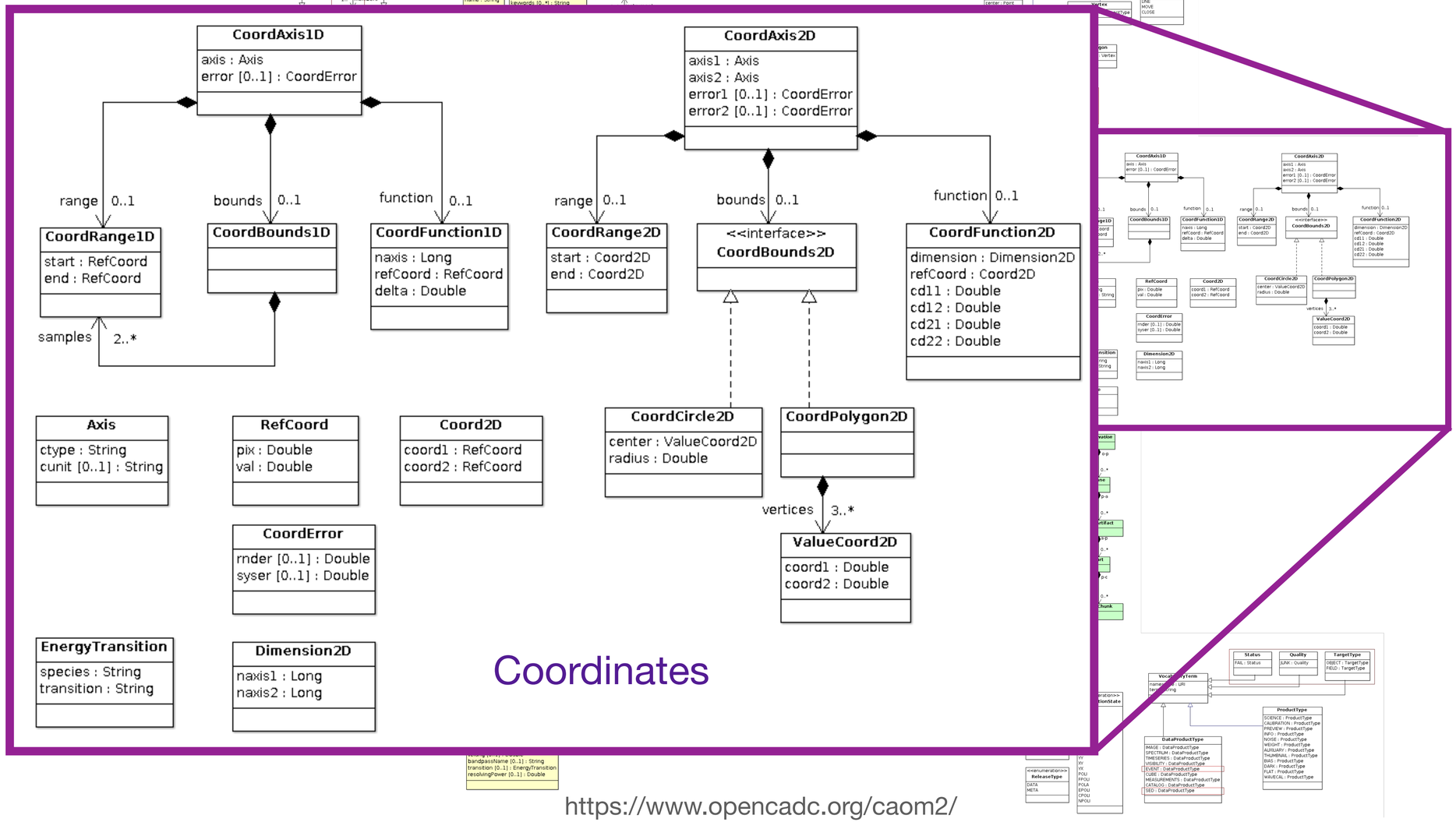
Overview



https://www.opencadc.org/caom2/

CAOM2 - Common Archive Observation Model

Overview



https://www.opencadc.org/caom2/

Dataset - CAOM Overlap

- Significant overlap in concepts between CAOM and Cube family of models with differences in organization and detail
- IF CAOM is eventually to become an IVOA standard, then we should make a plan for this while developing these models rather than after.

Goals of this discussion

- Dataset - CAOM
  - Is CAOM expected to become an IVOA Standard?
  - If so, or maybe even if not, how do we want to represent it w.r.t. the family of models which it overlaps?
- Dataset - Provenance
  - Determine approach for representing Observation Provenance in the Dataset model

Goals of this discussion

- Dataset - Characterization
  - When/How/Who do we revive the project to convert it?
  - Determine how/when this fits into the current model ecosystem.
    - Is it something which describes the current content of the dataset? Or more of a summary of the observation parameters (domain) that the dataset resides in?
    - Is it useful within actual datasets? Or more for dataset discovery?

Topics in order of priority (to me)

- What is the plan for CAOM? Timetable? And how do they see that happening? Ie: What impact on the current/upcoming model work?
- Opinions on representing Provenance information in Dataset..
- Relation of Observation and Dataset (Extraction or Reference)
- Roadmap for Characterization
  - Is this ObsCore (Data Discovery) content? Or metadata transported with a Dataset instance? Or both?