NRC·CNRC

Common Archive Observation Model

Patrick Dowler
Canadian Astronomy Data Centre

Data Models - Tues May 14

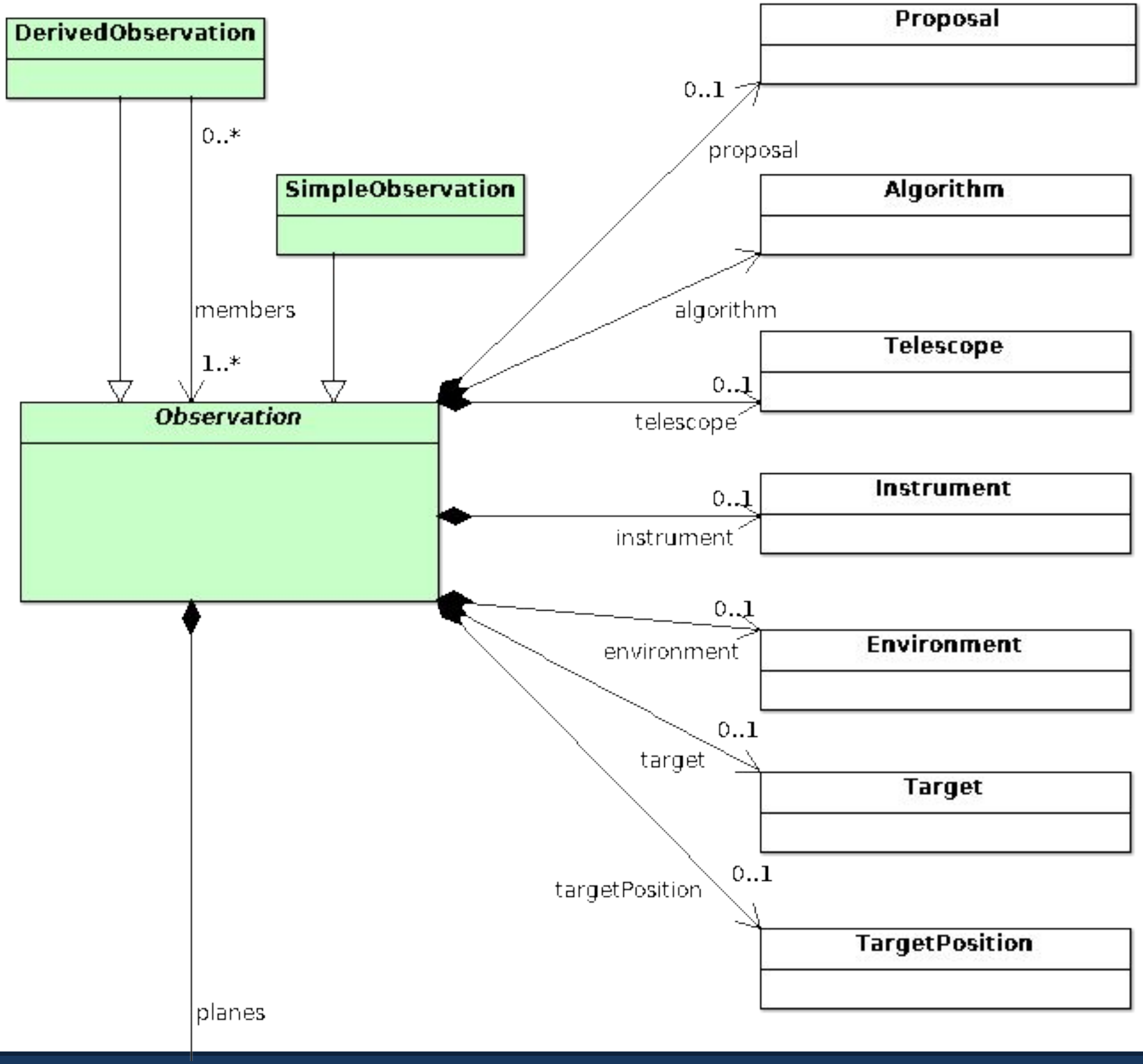
Common Archive Observation Model

- history, interest, and usage in data centres
- high level overview of CAOM
  - support for VO data models and APIs
- use cases that drive CAOM
- evolution of CAOM

Common Archive Observation Model - History

- CAOM-1.0 ~ 2007
- CAOM-2.0 ~2011
  - major lessons learned, implemented at CADC for all collections
- CAOM-2.2 ~ 2015
  - adopted at MAST for ~all collections
  - first CAOM workshop (2016)
  - remote metadata harvesting for HST Archive 2.0 project
- CAOM-2.3 ~2017
  - second CAOM workshop (2019): CADC, DRAO, EAO, ESA, STScI, INAF, IPAC, LSST, TMT
- CAOM-2.4 ~summer 2019

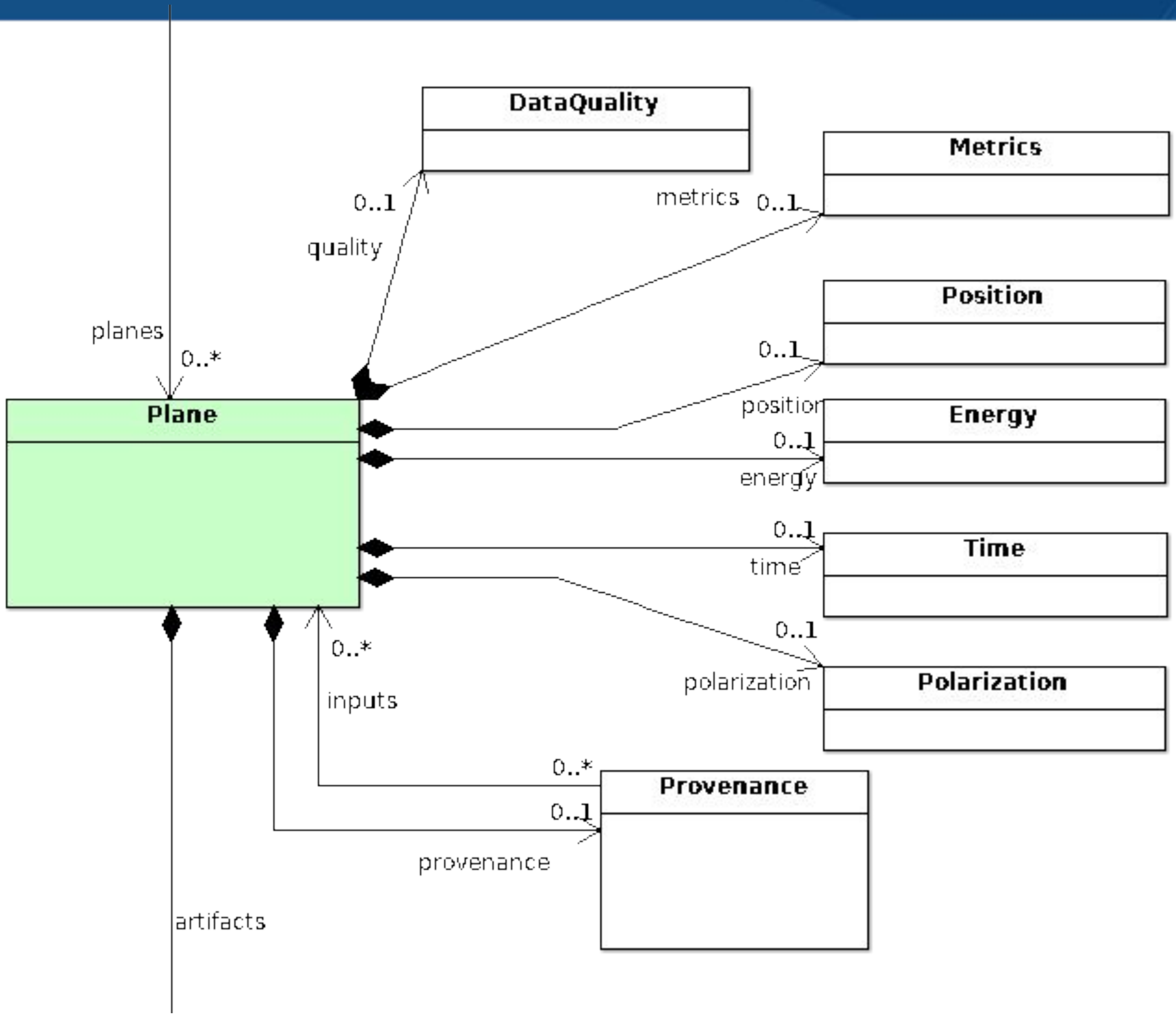
Common Archive Observation Model



#denormalised

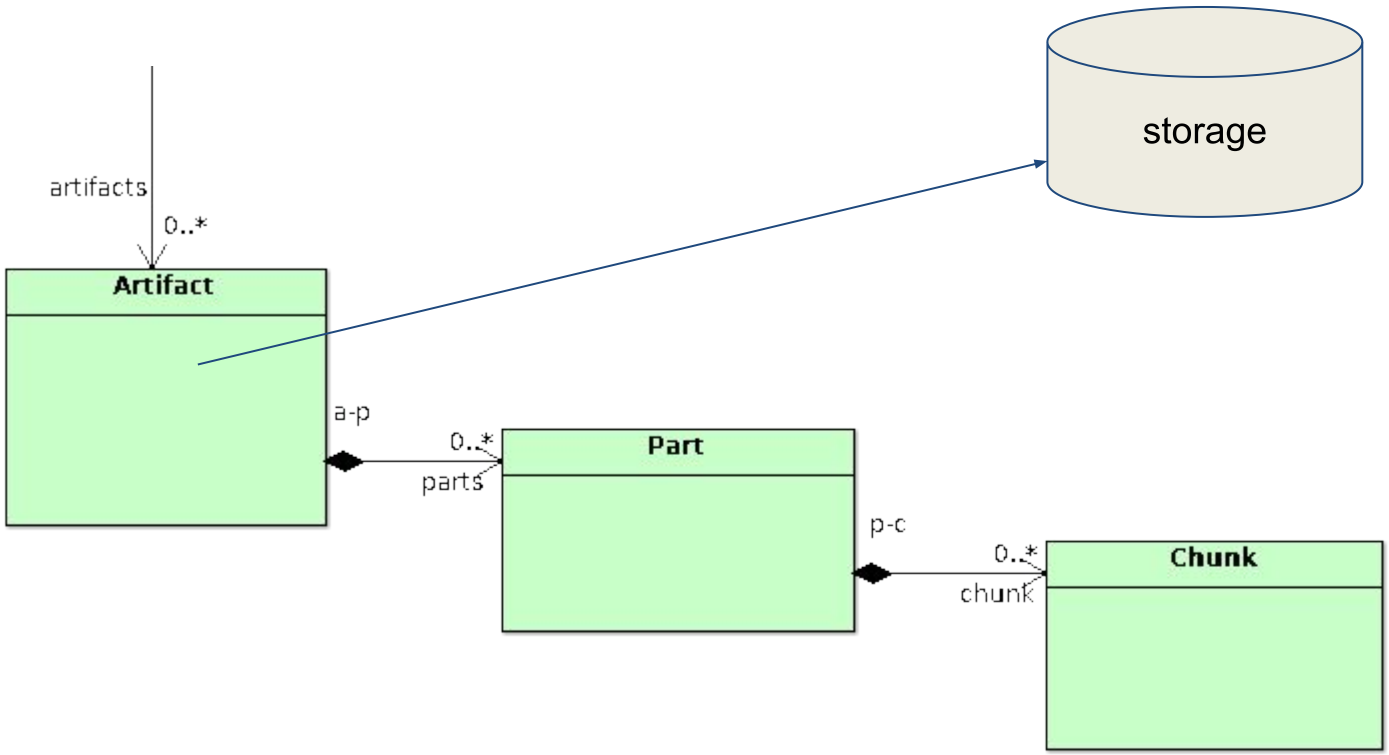
#descriptive

Common Archive Observation Model



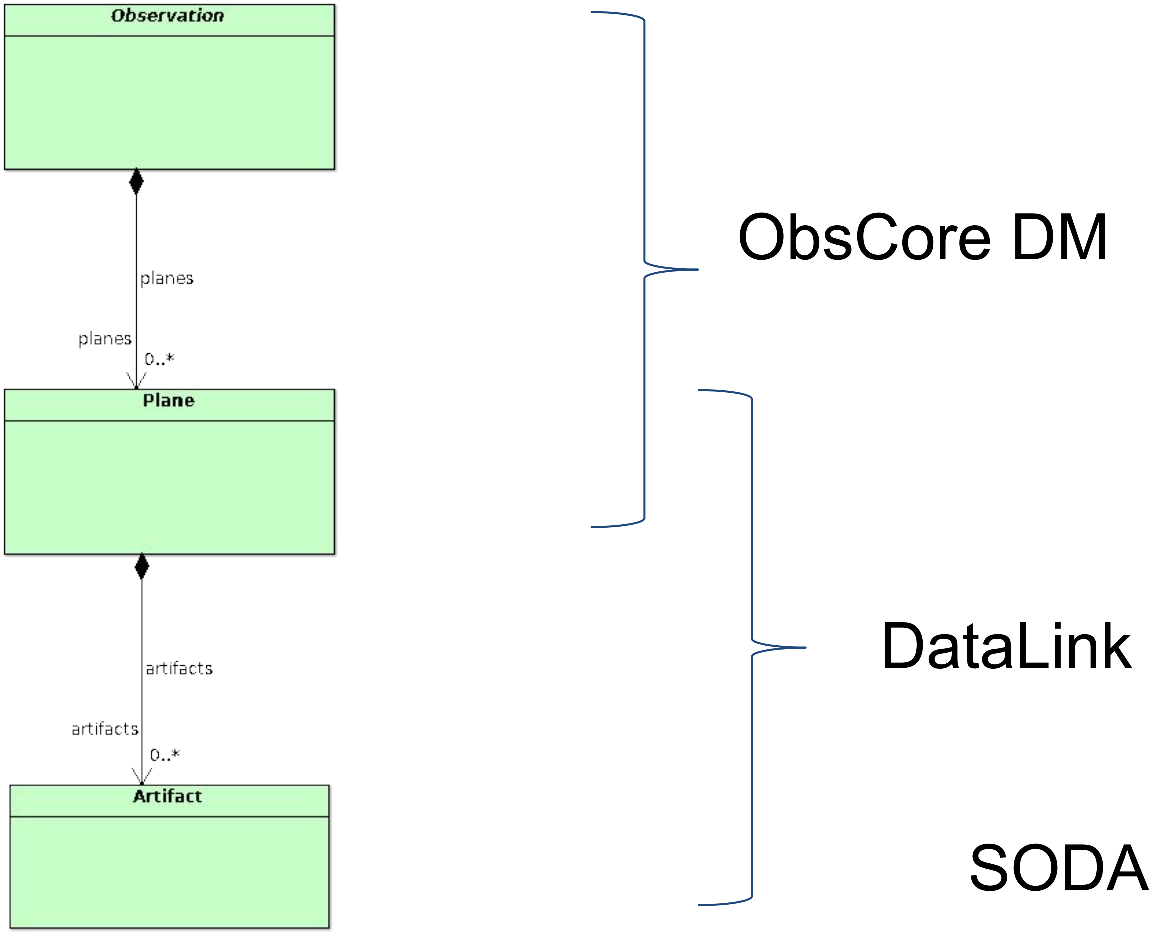
#product

Common Archive Observation Model



Common Archive Observation Model

- data discovery and data access & related IVOA standards



ObsCore DM

DataLink

SODA

Common Archive Observation Model

- **primary use case: describe the data we have**
  - with enough detail to differentiate in data discovery
  - with enough detail to drive data access
  - metadata needed to process/analyse: out of scope
- intended usage
  - static collections
  - data flow from telescopes → new instances available for discovery and access ~immediately
- new kind of data?
  - describe with existing model
  - evolve the model as necessary

Common Archive Observation Model

- **operational use case #1: computed Plane metadata**
  - data providers (telescopes) include complete WCS metadata in Part(s) and Chunk(s)
  - archive metadata service computes Plane metadata
  - Plane metadata was experimental / volatile
  - Plane metadata benefits from uniform computation / QA

- now: support hybrid mode where some collections provide all metadata -- others rely on service computations

Common Archive Observation Model

- **operational use case #2: SODA implementation**
  - CADC storage system supports cfitsio-style pixel cutouts
  - DataLink service: Artifact.contentType + presence of WCS in Chunk(s) predicts ability to perform sky-to-pixel transformation
  - SODA service: use WCS in Chunk(s) to transform user-supplied sky cutout (circle, polygon) to pixel cutout and generate a suitable storage system access URL
  - CADC storage system performs cutout-on-the-fly (currently)

Common Archive Observation Model

- **operational use case #3: incremental harvest of metadata**
  - support database-agnostic harvest of CAOM observation instances
    - database → database (redundancy, migration)
    - service → database (remote mirrors, sharing) *new*
- Observation.maxLastModified: timestamp maintained by origin server, used and copied by harvester(s)
- Observation.accMetaChecksum: stable metadata checksum a to verify correct serialisation and persistence

Common Archive Observation Model

- **operational use case #3: incremental harvest of metadata**
  - HST and TESS collections: MAST → CADC
  - HST collection: MAST → ESAC
  - HSTHLA collection: CADC → ESAC
  - operating with a harvest latency of ~5 minutes
  - full validation of a collection takes ~few minutes (HST: 1.5e6)
    - missed observations
    - missed deletions
    - mismatched metadata checksums
- currently harvesting whole collections -- other policies feasible

Common Archive Observation Model

- **operational use case #4: incremental synchronisation of data**
  - mirror (redundancy)
  - locate data near (in) processing resources
  - Artifact metadata used to figure out if download needed
  - full validation of a collection vs storage takes 10s of minutes (HST: 17e6 files)
    - missing files
    - orphaned files
    - mismatched checksums (+other file metadata)
- currently harvesting whole collections -- other policies feasible

Common Archive Observation Model

- **data model evolution**
  - current operation version: CAOM-2.3
  - under development: CAOM-2.4 with ~ 5-10 new features
- adopted a strict criteria for minor versions
  - **changes do not invalidate current metadata checksums** or otherwise require re-creation and re-ingestion of instances
  - instances are forward-compatible: new s/w can read old versions -- old s/w may fail to read new instances
  - operate in a hybrid environment with 2 adjacent versions

Common Archive Observation Model

- **data model evolution: what can change?**
  - add new optional fields (can be null)
  - change cardinality from 0..1 to 0..*
  - rename classes
  - rename fields: sometimes
  - change enum to vocabulary (with care)
- basically: don’t change the order that the metadata checksum algorithm accumulates bytes
  - proposed changes easily tested
  - data model libraries expected to support transitional API

Common Archive Observation Model - Summary

- CAOM is a data model with 10+ years of operational experience and evolution
- CAOM is in use in 3 multi-collection data centres for many data collections
- enables data discovery & data access
- enables metadata & data sharing… opens possibilities