

## DATA MODEL

Like text, published data files should be:

- Easy to find
- Easy to understand
- Easy to relate to each other

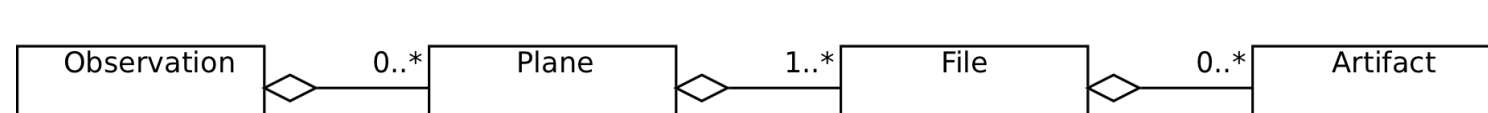
Headers should be based on a data model that captures:

- Physical properties
- Organizational properties
- Data Processing

CAOM (Common Archive Object Model)

is in use at the

CADC (Canadian Astronomy Data Center)



The CAOM backbone has four parts:

**OBSERVATION:** groups files based on observed photons

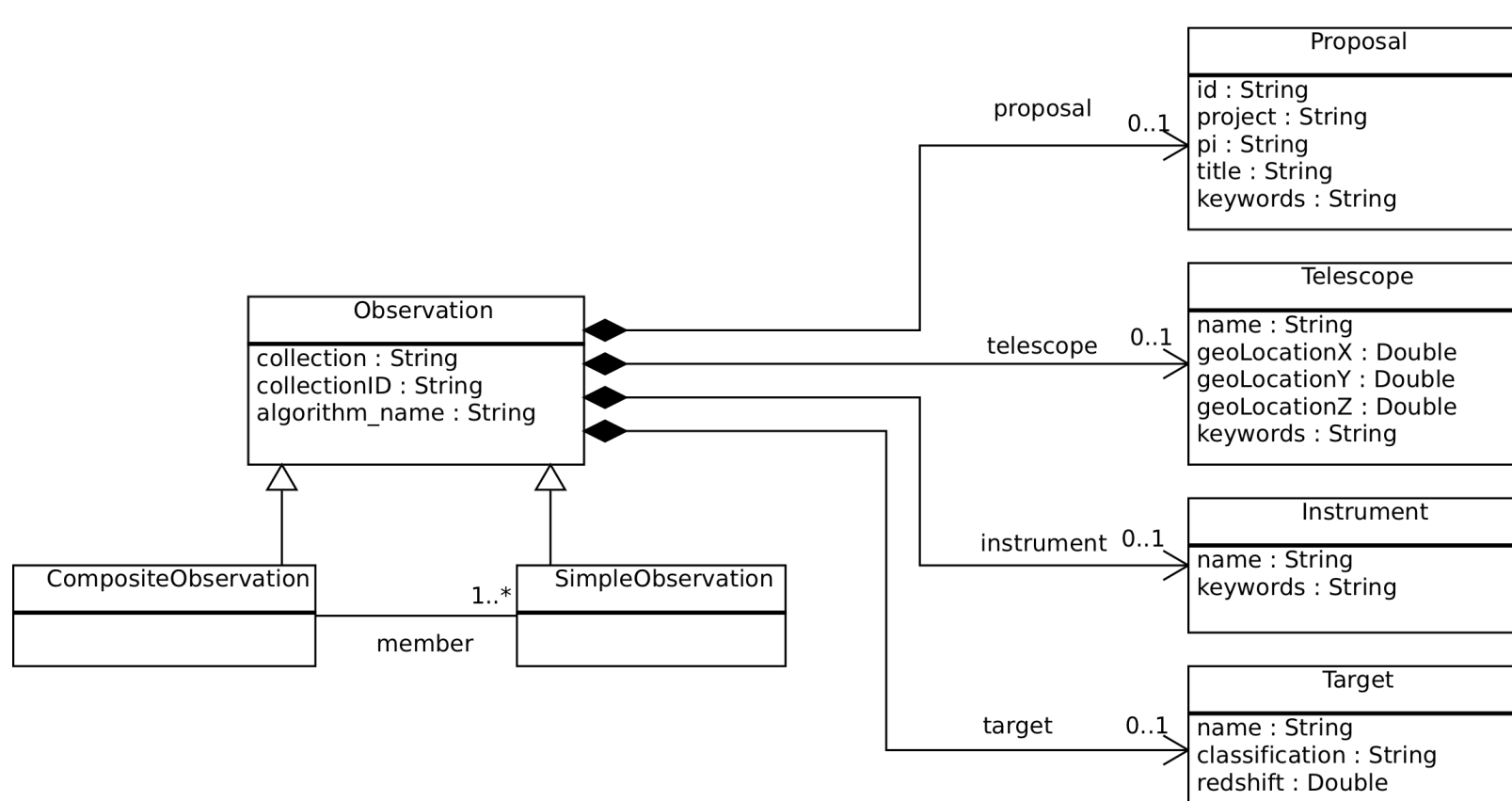
**PLANE:** groups files so closely related that they that would normally be downloaded together

**FILE:** The basic unit of data storage

**ARTIFACT:** (think pottery and tools) constructs within a file that can be described as arrays or intervals within a World Coordinate System (WCS)

## OBSERVATION

- Groups files based on observed photons
  - Simple – usually contain raw data from single observatory-defined observation
  - Composite – derived from data in a set of simple observations (members)



Uniquely identified by

- collection
  - usually telescope acronym
  - suggested header COLLECT
- collectionID
  - must be supplied by data provider
  - suggested header COLL-ID

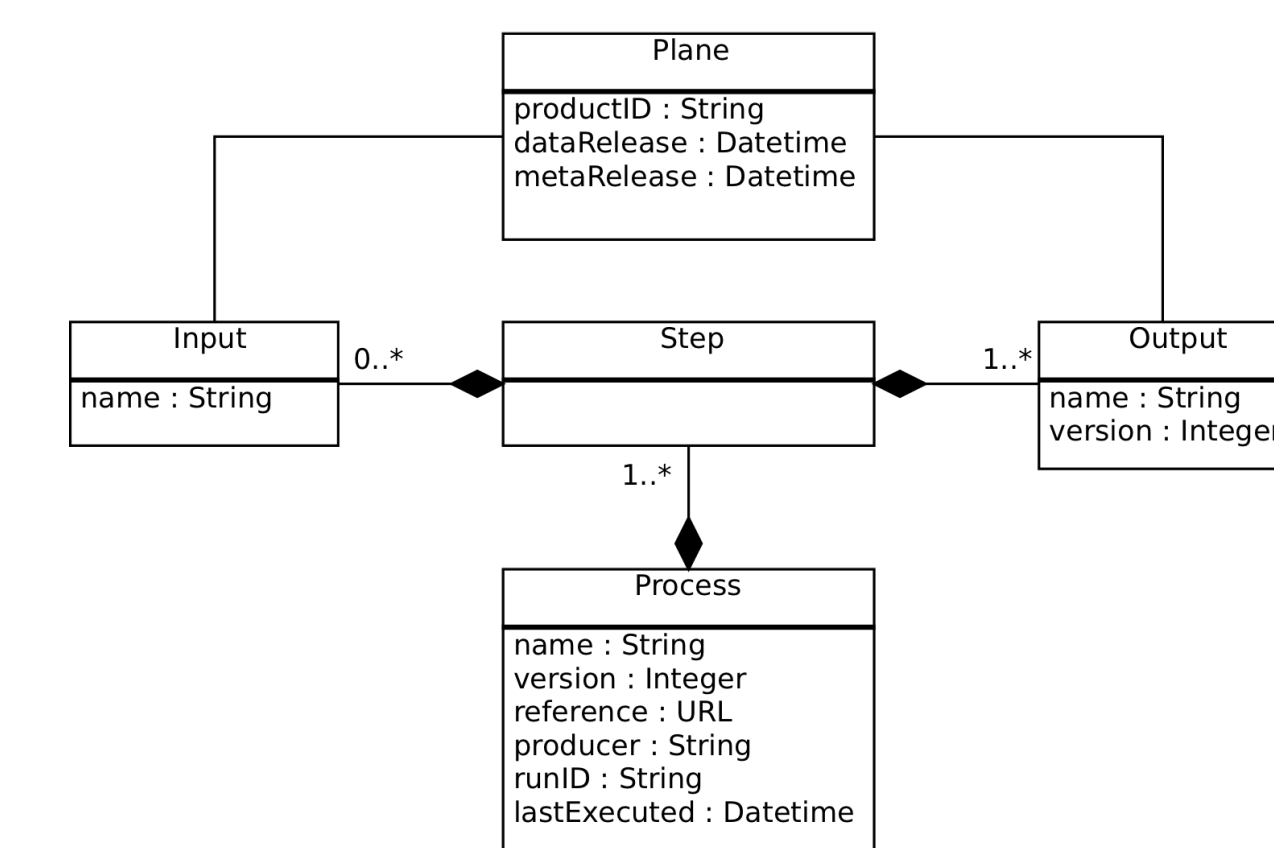
Fill with metadata for:

- Proposal
- Telescope
- Instrument
- Target
- Membership

## PLANE and PROVENANCE

- Plane groups closely related files
- Provenance records relations amongst planes derived from data reduction

- Process – data processing run
- Step – individual step within the run
- Input - role (name)
- Output – product (name) and version



Plane uniquely identified by

- productID
  - value/algorithm supplied by data provider

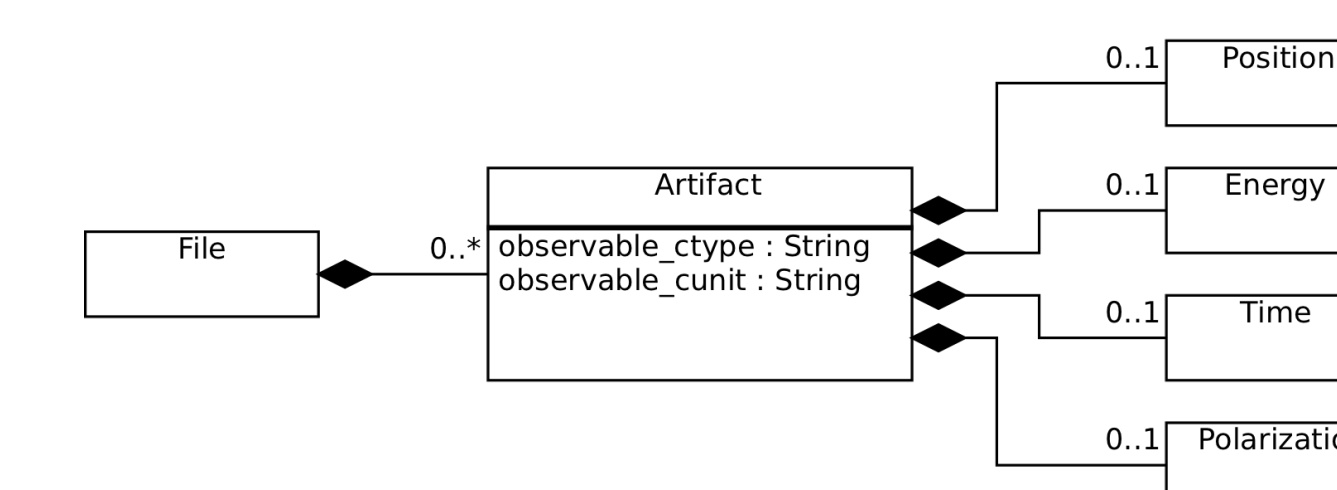
Provenance indicated in file headers using input file names

- PRVCNT
  - number of input files
- PRV1, ...
  - names of first input files

## FILES and ARTIFACTS

Files are the basic unit of data storage

- normally download whole files
  - can can do cutouts
- complex structure described by artifacts



File is a container class

Artifacts carry coordinate information

- FITS-style WCS for arrays
- intervals or bounds in some WCS

Fill with metadata for

- position
- photon energy (wavelength or frequency)
- time
- polarization

## OTHER ISSUES

### Output Products

- specify complete list
- input/output relations
- siblings

### Version Numbers

- sequential or date-based
- when to increment
- allow/forbid file replacement

### File Names

- unique in archive
- combine collection, collectionID, productID, etc.?

### Validity Checking

- tools like fitsverify
- warnings are errors
- mandatory header list?
- permitted header list?
- checksums

### Raw Data

- archive or omit?
- least-processed alternatives?
- source of metadata?

### Input Roles

- specify complete list

### Generation of productID

- from header? (PROD-ID?)
- using algorithm?
  - specify needed headers

### Order of Ingestion

- input/output relations
- dependencies for files without metadata

### Previews

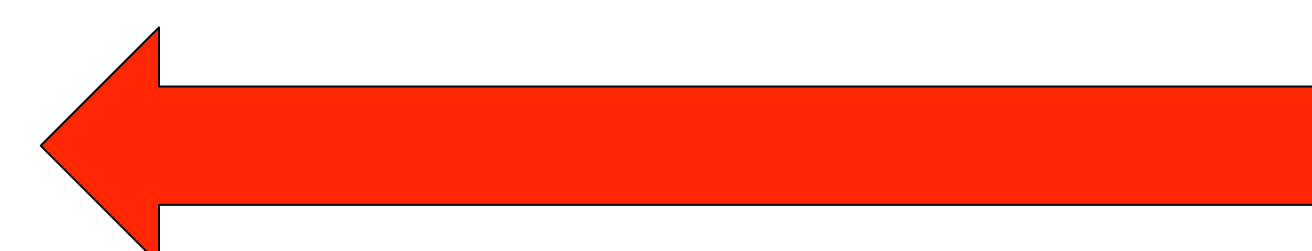
- generate during data reduction?
- generate with archive-supplied software?
- names and metadata

### Ancillary Data

- names
- dependencies
- source of metadata?

## ADDITIONAL INFORMATION

A more complete discussion of this topic is in preparation for publication. A draft can be found at: [http://ftp.hia.nrc.ca/pub/users/ror/ADASS\\_2010/File\\_Preparation.pdf](http://ftp.hia.nrc.ca/pub/users/ror/ADASS_2010/File_Preparation.pdf)



A full specification of CAOM version 2 is also in preparation: Patrick Dowler et al., 2011, "The Common Archive Object Model", (in preparation)