Data Model for Observation

Version 0.23

IVOA DM WG Internal Draft 2004-05-16

Working Group: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel
This version: Added Alberto's introduction.
Editors:
Jonathan McDowell, Fran cois Bonnarel, David Giaretta, Gerard Lemson, Mireille Louys, Alberto Micol
Authors:
IVOA Data Model Working Group

Abstract

This document defines the high level Observation data model giving the structure of metadata used to describe the content and context of datasets.

Status of this document

This is a Working Group Internal Draft only. It is inappropriate to reference this document.

Acknowledgements

Members of the IVOA Data Model Working Group, including representatives of the US NVO, Astrogrid, Starlink, and the AVO have contributed to the present draft.

Contents

1 Introduction and Scope

This document defines an abstract data model called Observation. In the early sections of the document we introduce concepts and discuss their interaction. In the section called ‘Model’ we present a formal UML class model using the concepts defined earlier. Serialization of the model as XML is deferred to a separate document.

1.1 What is an Observation DM?

An Observation DM is a way of describing an astronomical observation, and provides a way to associate with it things like (in no particular order):

- Observation coverage (in arbitrary axes, e.g. spatial, temporal, and spectral)
- Observation sampling (how it is digitized in the coverage axes)
- Observation accuracy (e.g. astrometric, photometric, temporal)
- Observation reduction and calibration (what processing has been applied to the data, and what calibrations were or should be applied)
- Observation mosaic (if the product is stitched together from several original datasets either as a spatial mosaic, spectral energy distribution or other combination, the recipe that was applied to do so)
- Observation circumstances: the telescope, instrument, detector and mode used including the various optical elements such as filters and gratings.
- Observation publications: links to bibliographic resources explaining the data product or scientific results from it.
- Observation plan: what was intended, by who, why, observing requirements and planning
- Observation packaging: how are the actual data packaged in files and data structures within the files, compression, etc.

The standard Observation DM is a way to provide an interoperable description of an observation. It is a conceptual, logical way to describe, tag and group information related to an observation, and is not tied to the way data providers internally store, describe, manage or organize their archives. It is conceived as a set of rules to be applied when describing an observation, either known to exist in some data repository or possibly to be created on the fly as the result of a query. These rules become important when exchanging data and/or metadata in the VO environment; only by applying such standard rules can we ensure interoperability.

Although we cast the model in terms of a definite observation made by an astronomical instrument, the model also applies to observational data created by merging, massaging or otherwise processing other observations, or indeed to simulated observations; thus the reduction and mosaic elements above may describe how the data was generated from another observation, or from a simulation recipe.

We do not at present propose the Observation DM as a model for secondary extracted measurements such as source properties, although we expect the model for such data to reuse many of the same components. In this document we include, but do not define in detail an ObsData object to represent the measurements themselves; the model may be used in conjunction with such an object to actually hold these measurements, or just to describe them. When we refer to 'metadata' in this document, we mean any data associated with the observation except for the astronomical measurements themselves.

1.2 What is an Observation DM for?

- a standard Observation DM allows interoperable exchange of metadata
- standard queries and answers become possible
- with a standard description of data products, visualization, reprocessing (e.g. mosaicing) and analysis are facilitated.
- The standard model paves the way for the development of new VO-enabled tools.

1.3 How is an Observation DM used?

The Observation DM can be used in different ways depending on the context.

In the context of the DAL (IVOA Data Access Layer), the DM will provide standard tags to formulate a query to a VO-compliant data provider (the Coverage part of the model described below will play a frequent role here) and a standard to describe the results of such a query (like the metadata tree used in IDHA).

In the context of data processing and analysis, the DM will provide a standard way to describe the accuracy, the resolution and the sampling applied to any observation. This lets tools handle observations from different archives in a systematic way.

The description of the instrument configuration used to collect the data is useful in a variety of analysis and query contexts.

1.4 What common problems are addressed by a standard Observation DM?

Examples of common problems are:

- Characterization of complex instruments and data products: each instrument and each data product has its own peculiarities; to come up with a standard model to accommodate all needs is not an easy task, but we can provide a framework giving a place to put most or all of the relevant information.

  The idea is to describe things hierarchically in levels of growing complexity. The coverage part will be the most used (and simplest) level, abstracting away all instrument and processing details. Lower in the tree we find more complex parts relating to instrument characteristics, modes, detectors, throughput curves, calibration corrections, rebinning etc.

- Commensurability: when manipulating data, it is important to know when two similar quantities can be compared, or when a particular data product is a compatible input for a specific algorithm. Use of the DM can ensure that each quantity can be identified sufficiently precisely.

  The UCD attributes provide a controlled vocabulary for astronomical and physical concepts. These, or a more precise extension of these, combined with the data model element structure itself (or, when serialized in VOTABLE, the corresponding UTYPE attributes) allow us to unambiguously describe the data.

- Packaging: if the data are provided in FITS, we may often need to ask questions such as: where and how are the pixels stored? where are the errors stored? is data quality information provided? is the data from a multi-detector instrument (e.g. mosaic CCD)? If so, how are the data from the different detectors arranged in the file or files?

  [We have yet to work on the details of this part of the model.]

2 Observation class summary

The Observation class describes a single dataset which may be:

- A dataset corresponding to an observation of the sky,
- A dataset derived from many observations.

with the stipulation that the dataset is intended to be analysed independently of other datasets, and contains all the primary data needed for such analysis. This purposely somewhat vague escription (without specifying 'primary' or ‘analysis’) leaves it to the data provider or user analysis system to decide how best to carve up the available data into observations. The intent is that a single science exposure of an instrument (although even that can often be hard to define) will usually map to one dataset. This will often include multiple arrays of numerical data, such as images from several CCD chips.

The model offers a single self-consistent description of (1) all the metadata needed by data analysis applications, as well as (2) metadata needed for data selection and retrieval. Case (2) is usually a simplified version of the case (1) metadata. Early implementations of the model will emphasize this simpler use case with the guarantee that the underlying model is already extensible to the fuller description.

An Observation can be a spectrum, an image, a time series, or a higher dimensional combination of these, as well as an interferometric visibility dataset or a photon event list, or a source catalog generated from a single observation (such as a single sextractor run).

We do not discuss theoretical (simulated) data here, but we anticipate that simulations which are intended to directly model observational data will be represented by a Simulation object which is simply an Observation object in which the the provenance discussed below includes metadata describing the simulation process.

The Observation model will be used in two ways: (1) we provide a serialization to XML which data providers can make use of to describe their data in a standard way, and (2) we forsee that registeries, data access and analysis services will operate on the software objects defined by the Observation model.

Data providers will describe their data with different levels of completeness: some will offer only simplified descriptions suitable for searches while others will support detailed descriptions, and yet others will include full traceability to earlier levels of processing. We expect to describe this by defining ‘levels of compliance’ with the observation model (a level 1 compliant description would include only the simplified metadata, etc.).

2.1 Use cases

Some use case datasets represented by the Observation model include

- A 2D sky image with an RA,Dec coordinate system and metadata describing its coverage
- A multi-detector image (such as a mosaic imager, in which we retain information about which parts of the sky are covered by the individual chips)
- A mosaic image, in which different misaligned fields, possibly from different instruments, are coadded to create a larger image with a complicated sky coverage; the difficult thing here is to retain the multiplicity

of instrument information. In a coadded mosaic, we do not retain direct information about which regridded pixel has contributions from which original detectors, although we may be able to reconstruct such information from the metadata.

- A 1D spectrum
- A velocity cube
- An X-ray event list

The actual use cases are:

- Search for these datasets in an archive by a query involving their metadata
- Describe the interface to these data as part of an interface between interoperable tools.
- Display the hierarchical nature of an archive, including user-selected subsets of a dataset.

2.2 Relation to Quantity model

The Observation model for astronomical metadata is built on top of the Quanitity model which models pieces of data and metadata. Quantity associates single values or arrays of values with a UCD and unit, and supports describing the axes of an array and the coordinate mappings on those axes. It can therefore be used to describe the data within an observation, as well as the individual metadata associated with the observation.

3 Observation Components

We define the Observation as an aggregation of components, discussed individually here.

3.1 Observation Data and Observable

The Quantity object is used to contain the main data for the observation (although it may be used in other contexts too). This object is described in a separate document; here we illustrate its use in Observation Data, a trivial subclass of Quantity.

Often the Observation Data is a single, scalar Quantity - often but not always a photon flux of some kind. The Observation Data may be a compound type made of several Quantity objects. An x-ray event list typically consists of a Photon quantity compounded from individual time, spatial coordinate, energy and other quantities, together with a derived Flux quantity defined by an integral mapping on the Photons. Another common compound case is the pairing of a Source quantity with a Background quantity, sharing the same observation metadata.

One of the Quantities in the Observation Data, usually a non-compound Quantity, is always designated as the Observable and represents the data values of primary interest such as flux, luminosity, etc.

3.2 Curation

Registries may harvest Curation metadata from individual datasets. As a matter of consistency, Observations should support Curation metadata as specified by the Resource and Service Metadata data model. This will ensure traceability of datasets.

What Registry describes as a matter of Curation items is a high level, general purpose description at the level of a mission, a survey, an experiment. As such general metadata we shall find Institution name and adress, PI, name and adress, distribution service name and adress and contact info. At the level of each individual Observation, we may support more detailed curation information. In particular the curation should mention the data reduction pipeline and the appropriate contact information for the data reduction and calibration, in addition to the conventional observer information. Another component of the curation, describing the context of the individual observation in a larger context, is the ObservingProgram class. This includes information about surveys or large-scale observing proposals of which the observation forms a part. The early RSM Curation model will be extended within Observation to incorporate these details.

3.3 Characterization

Data characterization is the meat of the observation model. It includes Coverage, Sensitivity and Resolution.

We emphasize in our model the relationship between different levels of simplification. In general, data query/discovery use cases will tend to use simplified representations, and data analysis/tool interface use cases will use the most complete representations.

The table below shows different kinds of characterization used in astronomy in the spatial, temporal and spectral domains and in the observable dimension (assumed to be flux). We use these examples to motivat the definitions of our general characterization concepts: Location, SensitivityBounds, Support, SensitivityFunction, Filling Factor, Resolution and Sampling Precision.

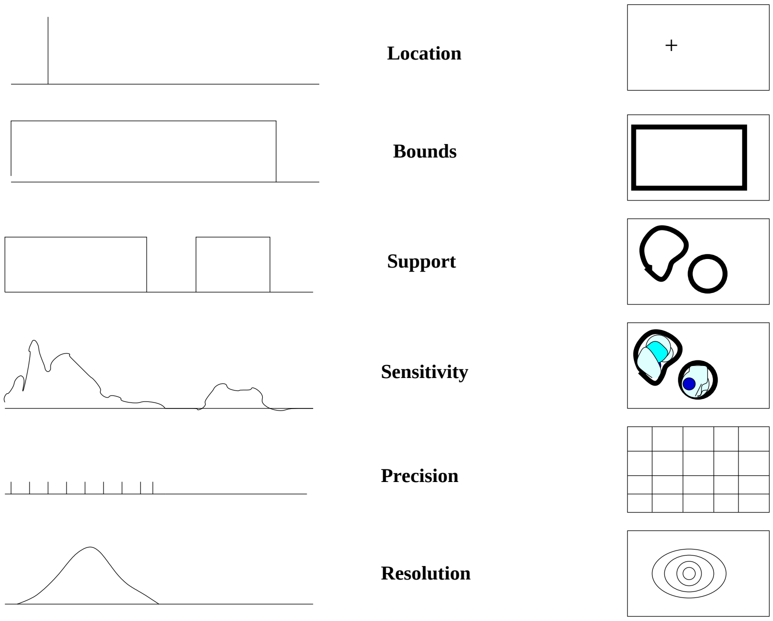

Figure 1: 1 and 2 dimensional illustrations of various aspects of data characterization.

General	Spatial	Temporal	Spectral	Flux/Observable
Location	Pointing direction	Observation date	Band	-
SensitivityBounds	-	Start/stop time	Bandpass	Saturation, pileup Limiting flux
Support	Field of view	Window function, Good time intervals	Bandpass	-
SensitivityFunction	Flat field, QE Vignetting		Transmission curve, Effective area	Linearity
Filling factor	Pixel f. fac.	Dead time		
Resolution	PSF Res'n (FWHM)	Exposure time	Line spread function	-
Sampling Precision	Pixel size	Exposure time	Channel width	Digitization
Mapping	WCS	Clock rate	Spectral WCS	Zero point Counts/flux func

The concepts are not always fully separable. For example, high energy missions move the telescope during the observation, leading to a time-variable mapping from detector to celestial coordinates called the ‘aspect solution’. This in turn leads to effects such as spatially variable effective exposure time. All of these concepts are well-defined for other domains, but don’t always have domain-specific names. It’s a little harder to see the equivalents for theory parameters. The simulation grid gives a sampling precision, but resolution may be hard to determine without post-processing.

The Characterization object shall be extensible so that other specialized parameters (e.g. ambient conditions such as airmass, or instrument operating temperature) can be described by them, although at this stage we may not define the representation of such parameters.

3.4 Coverage and SensitivityFunction

We consider several different levels of description of the depth of an observation. The most simple description is a Location in arbitrary parameter space - for example, the statement that an image is at a particular RA, Dec, and taken at a particular wavelength and time. Here the interpretation is that the values given are fiducial values representative of the data, with no precise definition (mean, weighted median, etc.) being required.

The next level of description is the SensitivityBounds, where we give a single range in each parameter. The interpretation is that all the valid data is guaranteed not to lie outside these bounds, but there may still be some values within the bounds for which there is no valid data. There is a slight loophole here with the word 'valid': for instance, if a spectral filter has a red leak, we may consider the frequency ranges in the red leak to be invalid data (with quality values so marked) and outside the bounds. This satisfies the intent of typical queries, which want to find observations which may have useful data within a given range of interest.

The Support component describes the more detailed context of the observation in a quantitative way. It will describe the space, time, frequency and other ranges covered by the data. Mathematically, the support of a function is the subset of its domain where the function is non-zero. In our model, we will fudge this slightly to mean the subsets of the domain where there are valid data (according to some specified quality criterion).

Note that these ranges may include the independent variables of the observational data samples as well as variables which are the same for each sample; thus for a 1-dimensional slit spectrum, the frequency range extremes of the spectrum (independent variable) as well as the time of observation start/stop and the region of the sky covered by the slit aperture (constants for the observation) will be described by the coverage. The coverage may have multiple ranges for a given parameter - particularly useful in the case of times, where an observation may consist of the co-addition of several widely separated time ranges. For two-dimensional parameters such as sky position, the coverage can be described by Regions (whose interface is described separately).

The most detailed description of the depth is Relative SensitivityFunction, which goes beyond the on/off coverage description to a description of the relative cell-to-cell sensitivity in the data. This includes filter transmission curves, flat fields, sensitivity maps, etc.

However, in practice we do not use Support and Relative SensitivityFunction to describe the case in which there are a large number of small interruptions to the data. This arises in the temporal domain with detectors which have dead time between each sample, or in the spatial domain with pixels with gaps between them so that the active area does not completely fill the focal plane. In these cases analysis systems always handle the problem with a statistical correction, correcting the effective sensitivity by a Fill Factor (usually constant for an observation but sometimes varying with the coordinates).

The final level of characterization in this sequence is Absolute Sensitivity, which includes the upper limit value of the Observable (e.g. limiting magnitude) at a given position, and the value corresponding to one detector count in cases where that concept applies.

The values of these Coverage and SensitivityFunction Characterizations of an Observation may be derived from a number of factors, some of them described by other Characterization features and others by Provenance details (for example, the spectral sensitivity may have been derived from the Instrument and Filter in the Provenanc). These links between various attributes in the model will be reflected in the data model (using for example ”attribute formulae”).

3.4.1 Space-Time Coordinates

Arnold Rots has developed a metadata scheme for Space-Time Coordinates (http://hea-www.harvard.edu/∼arots/nvometa). The latest version of this proposal is actually a general coordinate specification, with space-time coordinates as a special case. We propose that the Location/Support elements of our characterization can incorporate the STC metadata.

We discussed extensively the choice between putting Resolution and PixelSize together with the Name/Value/Error in the Quantity object, which Arnold has strongly argued for, and putting them in the Characterization. The majority favored the latter approach so that the Characterization object would have all information relevant to discovery-type use cases, and to prevent the Quantity object from becoming too heavy. We may need to revisit this design choice.

With this proviso, we can construct a Coverage object which consists of an arbitrary number of axes. Some of these axes will be the same as the axes of the main Observation Data, while others will represent phenomena that have been integrated over. For example, the simple 2D sky image has celestial coordinate axes, but has also been observed over a finite integration time and wavelength band. The time and spectral axes are not present in the main data array, but their bounds - and even, for such things as color corrections, their sensitivity as a function of the coordinate within the bounds - may be represented in the Coverage.

The STC CoordSys object can be used to define the axes of the Coverage. Each CoordSys object will have a corresponding CoordArea object which gives the Support.

Since the space, time, spectral axes are particularly important for astronomy, we may wish to verify that a complete and consistent space-time-spectral description is present. We recommend that implementations include a method to return an STC AstroCoordSys object to provide this checking; an incomplete description will not be able to return one.

The STC definition also emphasizes the need to know the space-time coordinates of the observatory (actually the aperture), potentially as a function of time. We need to model this in the context of Observation - the observatory location will be part of the Provenance, but we will also need the space-time coordinates of the 'effective aperture' in the Characterization. This will be the observatory location for raw data, or the barycenter location for barycenter-corrected data, etc.

3.5 Resolution and Sampling Precision

The concepts of resolution and sampling precision (or pixelization) are related. Ultimately resolution describes the continuous smearing of our knowledge about the data, or more precisely the probability that a photon (or other observable) which has one set of attributes is measured as having a different set of attributes. Mathematically, if the physical attributes (e.g. position, time, energy) of the photons are $\mathbf{x}$ (e.g. x_0 = energy, x_1 = RA, x_2 = Dec, x_3 = time, etc.), and the measured attributes are $\mathbf{y}$ (e.g. y_1 = spectral channel, y_2, y_3 = pixel position, y_4 = time bin) then given a flux of photons $S(\mathbf{x})$ the detected number of photons is

$$N(y_1, y_2, ...) = N(\mathbf{y}) = \int \mathbf{S(x)A(x)R(x,y)dx}$$

where A is the probability that a photon is detected at all (the quantum efficiency) and $R(x_1, x_2, ..., y_1, y_2, ...)$ is the smearing of measured values (PSF, line spread function, etc.).

In the most detailed case, the R function may be specified as a function of the coordinates - for instance, a PSF which varies as a function of detector position and energy. The first level of simplification is to specify a single function which applies to the whole observation - e.g. a single PSF. This function may either be provided as a parameterized predefined function (e.g gaussian) or as an array. The final level of simplification is to give a single number characterizing the resolution, effectively implying a single-parameter default predefined function. We may support several versions of these simple resolution parameters; we propose initially that a resolution interpreted as the standard deviation of a gaussian be supported.

Sampling (or pixelization or precision or quantization) describes the truncation of data values as part of the data acquisition or data processing. If the sampling precision is small compared to the resolution, the knowledge of a single data value is limited by the resolution. If the sampling precision is coarse compared to the resolution, knowledge of a single data value is limited by the sampling. If the mapping of the data coordinates (the pixelized/truncated ones) to the coordinate axes is nonlinear, the sampling precision varies from sample to sample; the next simplification level is the definition of a 'characteristic sampling precision' for the whole observation.

The distinction between continuous smearing (resolution) and discrete quantization (sampling) often - but not always - reflects a physical distinction between the atmosphere/telescope optics combination and the discrete pixels and A/D signal conversion of a detector. More importantly for our model, it reflects aspects of the data which are handled differently in downstream data analysis.

3.6 Background Model

The term 'background' is used in two different but related senses. As all Escher fans know, background can easily become foreground with a shift of perspective. In general, 'background' is whatever signal (and noise) is left over after 'sources' have been removed, where the definition of sources is made by the astronomer (and may change each time the same observation is analysed). This can include background signals from the instrument, such as cosmic ray hits and bad electronics, and background signals from the sky from unresolved sources and from diffuse emission, including cosmological backgrounds and, sometimes, extended astronomical sources (for instance when detecting a star cluster inside another galaxy).

A second sense of background is a specific, designated subset of the observation, or another designated observation, associated with a particular source and deemed to be a **realization of the background model for that source**. Suppose we identify a source as lying in a 10 arsecond circular region of an image. To evaluate the flux of the source, we wish to model the contribution of the background (in the first sense) to that 10 arcsecond circle and subtract it. Often we do this by selecting a much larger (to suppress counting statistical noise) region in the same or a similar observation which is deemed to have equivalent statistical properties, and scale it to the appropriate area. We don't just take all the area of the image not tagged as source areas, because the detector properties may be variable and the background may vary across the image. We loosely refer to this dataset used to model the background for a specific source as simply 'the background' for that source. In that sense, the background belongs logically with the data for the source region (the total signal for the source region) - the paired datasets imply the net (source minus background) signal and some information about the errors on that signal. In many archives, particularly spectral ones, the source background and the source total are stored together on a common coordinate axis, and it makes sense to model the pair as a concept and put them in a single Observation.

3.7 Provenance

The Provenance is the description of how the dataset was created. For many analysis tasks, information about some aspect of the data acquisition chain is needed. The Provenance object provides a flexible structure to store such information when appropriate.

In an experiment which detects photons emitted from an astronomical source, we can loosely identify five phases of then observing process:

- **emission** of the photons by an astrophysical source,
- **propagation** of the photons along the line of sight to the observer,
- **observation** of the photons with a telescope or other instrument,
- **data processing** of the output of the instrument to an archived data product, and
- **analysis** of the archived data product to create an output product with a scientifically useful result.

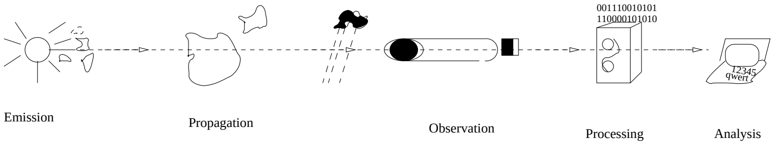

Figure 2: The photon's journey to the ApJ

These steps are meant principally for observational data but could also be used to describe the theoretical assumptions and configuration used for simulated data . However, the boundary between each of these phases is blurry and even for a given dataset may depend on the science question asked. For example:

- Is the VLA correlator part of the observation, or part of the data processing?
- Are telluric absorption lines part of the propagation and treated on the same basis as interstellar absorption, or handled together with instrumental effects and therefore though of as part of the observation?
- Is a gravitationally lensing cluster of galaxies part of the propagation, or part of the telescope?
- Are quasar broad lines to be considered part of the emission, or a propagation effect on the nuclear continuum?

These questions suggest that our model should not emphasize these boundaries too much, and treat each of the phases as part of one overall process. Nevertheless it may be useful to retain the distinctions. *The boundary between observation and data processing is perhaps the most useful one, and it should reflect the earliest point where the data provider would begin again if improved processing techniques became available, that would not involve detection of photons from the sky. Thus in the example above, the correlator would not be part of the data processing, as correlation would probably not be repeated.*

We also introduce the concept of the Proposal, which includes the formal observing proposal if any, the identity of the observer and the target, and the intended protocol for the observation. This is not always the same as the actual observing protocol - malfunctions, errors, and real-time decisions make the actual observation different from the proposal.

3.8 Instrument

One of the sub-objects of the Provenance that has been extensively modelled is the Instrument. The most important metadata associated with the Instrument are the location and direction of the telescope aperture; for ground based data this may be the latitude, longitude and height, which usually are constant in an Earth-fixed coordinate system on the timescale of an observation, while for space-based or aircraft/balloon data an ephemeris is required. Other metadata that may be important for analysis are those describing the ambient weather at the observatory (e.g. cloud cover, temperature, particle radiation flux, magnetic field strength). These metadata depend on the realisation of the observation and not on the device configuration. Therefore a class could be designed for that, named ObservingConditions or Ambient-Conditions.

We describe the Instrument in a hierarchical model consisting of an Observatory containing one or more Observing Configurations. An Observatory may have multiple telescopes sharing common metadata. Space observatories frequently have independent optical systems mounted on the same host satellite. Interferometric data will require special attention; optical interferometers combine light from telescopes which are physically close to one another and may share metadata on ambient conditions, while radio interferometers may combine radiation gathered by widely separated telescopes, even on opposite sides of the Earth or by combining data from Earth-based antennae with one or more antennae in space. These must be distinguished from proposed missions such as Constellation X, which will co-add light gained by widely spatially separated X-ray telescopes, using multiple optical elements not for interferometry but simply for increased total effective aperture. We propose a model in which the Observing Configuration is made up of an ordered list of Observing Elements, and we specialize the Observing Element class to Aperture, Optic, Grating, Filter, Camera and Detector classes. (It's the location and axis direction of the aperture we care about - even if the Detector is at the other end of a 1 AU optical path). Some observing configurations have multiple optical elements, filters and gratings, and there is no standard order for them. (However, putting the Detector before the Optic in the path will probably not result in optimum data.)

For many purposes these details may not be required, as they will be abstracted into parameters like the effective spatial resolution and the effective sensitivity of the final dataset. Those idealizations will be kept more directly attached to the data, while the Provenance can be used to record these messy details.

In particular, instrument noise parameters will be attached to the appropriate Observing Elements, while general abstracted noise characteristics will be associated with Accuracy objects in the Observation Data.

3.9 Calibration

Describing the calibration files associated with an observation is particularly problematic. We may consider several classes of calibration files: those taken specifically with a given observation; those belonging to a restricted set of observations (such as a superflat taken from the median flat field for a whole observing run); those applying to all observations taken in a particular configuration; and those that are theoretical or fixed by definition.

(Distinguish between calibration data and applying a calibration; calibration data as observations themselves).

In addition, when modelling archives and analysis systems we may need to distinguish between calibration data packaged with each observation, calibration data which are stored in a common area, and calibrations which are applied in code without an exterior calibration dataset. If a calibration dataset is the same for all HST WFPC-2 observations, a VO data request for 100 WFPC-2 images should not return 100 copies of the same calibration file - this puts a requirement on the DM to be able to handle such situations.

Calibration files are generally associated with either the values in the observable quantity or with the axes. Thus for a spatial image we may have flux calibrations, astrometric calibrations, wavelength calibrations, and so on. This is not enough to define a calibration - flux calibrations in this sense include bias, flat fields, linearity corrections, cosmic ray removal, and the conventional flux cal (counts to flux conversion).

In general we may define any calibration as a mapping of a quantity, which may change its UCD. One task for the UCD and DM groups to discuss is whether calibrations which merely correct one physical phenomenon (e.g. bias subtraction, dereddening) rather than changing the units (e.g. counts to flux) should merit separate UCDs. To put it another way, should the distinction between raw counts and bias corrected counts, or that between observed flux and dereddened flux, be described by the UCD system or by some other piece of metadata? The tentative conclusion of discussions in

Garching (Jan 2004) was that other metadata attributes should be used, rather than using the UCD.

The close association of most calibrations with particular axes or observables suggests that they should be modelled in conjunction with those Quantities. Calibrations are also closely associated with the Observing Elements - but particularly in ground-based work, an individual calibration may represent the net distorting effects of several Observing Elements which cannot be disentangled, so tying the calibrations to the instrumental configuration is often not practical.

We therefore propose a separate Calibrations object within the Characterization that can answer questions like 'is this flux dereddened' or 'is the wavelength scale corrected for geometric distortions' - these questions fall somewhat between 'what UCD?' and 'what coordinate frame?'.

3.10 Target/Field

Many query use cases involve looking for particular astronomical objects or classes of object in ways which a simple celestial position search can not satisfy. In particular, questions of hierarchy - 'give me all the Be stars which are members of the Orion OB association' - require an AstronomicalObject model and indeed a Universe model. Some analysis tasks also require knowledge of object properties - for example, corrections to the rest frame of the object.

We distinguish between:

- an astronomical object - which has properties independent of a particular observation, and has associations with other objects in the aggregation of the Universe -

- a source, which is an entity created by analysis of an observation (e.g. applying a source extractor to an image), representing the detection of an object (or alleged object) by an instrument - with an observed flux, etc.; the process of association of one or more sources with a single astronomical object is called *identification*, and is emphasized in the VO context because variability, confusion and finite spatial resolution, and other observational effects can easily cause erroneous identifications, so that a positional identification of, e.g., an X-ray source with an infrared source may or may not be reliable. This problem also arises for solar system objects; an asteroid provisional designation such as "1933 OB" corresponds to a source, while a final number designation such as (4589) corresponds to an object and typically results from identifying several provisional designations.

- a target, the thing the observation was made to study. Often the target is an astronomical object but sometimes it is a field.

- a field is a region of the sky (usually - but possibly a region on the surface of a planet or other object?) not corresponding to a physical object in 3-dimensional space, and possibly associated with a name. Usually we use celestial regions as simple anonymous subsets of the celestial sphere (see the work of A. Rots on regions) but, especially in the context of the Observation model, we can associate a permanent name with a region ("Selected Area 54", "Hubble Deep Field South", "Lockman Hole", "Antlia", "Zezas et. al. SMC Field 5"). Note that this is subtly different in usage from a region delimiting the sky extent of a physical object ("NGC 225"), where the contour is thought of as being associated with a distance, and hence a mapping from angular to physical scale.

4 Data Collection/Archive

In the AVO demo, an IDHA-based data model provides a very useful data exploration tree to allow exploration of an archive (which may be an arbitrary local data collection or a full-up data center) down to pieces of a single dataset. For the IVOA model we propose a separate DataTree model to support this, but the idea that this tree extends within an observation requires its discussion here.

The AVO/IDHA tree is layered (branched) by Observing Program, Object and Filter. We can generalize this concept by allowing a tree ordered by any specified set of metadata tags, allowing users and data providers to easily structure their desired view of a data collection. By providing methods to search a simple list of files to harvest the metadata, one could allow construction of such a tree even when no archive database is present.

The AVO/IDHA tree has sub-observation leaves of two kinds: observations where there are several detectors (multiple chip imagers), and observations where there are more than two dimensions (velocity cubes) and velocity slices can be seen as individual leaves. IDHA treats these as identical cases where an observation is made up of multiple pieces. However, the distinction between these cases is that the pieces in the first case are uniquely defined (the separation between chips is a hardware feature) while in the second case there are multiple ways one could slice the data (in an RA, Dec, V cube, it is equally legitimate and frequently desirable to look at a set of RA-velocity images along Dec slices instead of RA-Dec images along velocity slices). Another distinction is that in the first case the coordinate systems may be entirely independent from chip to chip (the pixel sizes may even be different) while in the second case a single 3-dimensional coordinate definition can describe the data (the same x,y maps to the same RA,Dec for different velocities).

We can support both these cases by defining methods which define a new Observation as a subset of another one: we should be able to ask for the number of pieces of an observation (multiple chip case) and we should be able to define on the fly the splitting of an observation into pieces (data cube case).

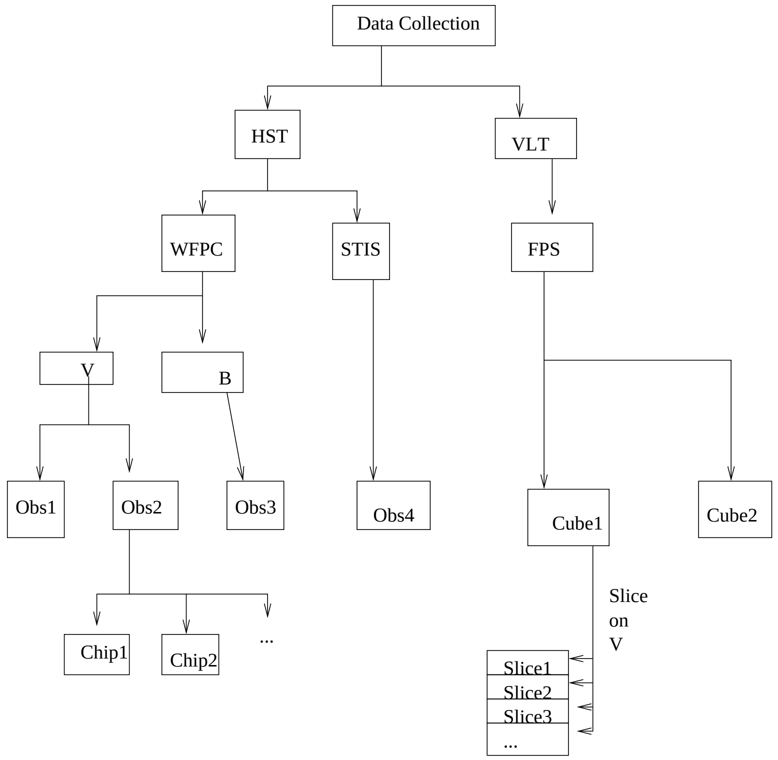

Figure 3: A possible instance of a Data Tree, showing multiple-chip and velocity-slice leaves.

5 Relationship to Source Catalogs

We will model catalogs in a separate document. We note here that for a simple source catalog derived from one Observation (e.g. via Sextractor) the model presented here can be associated with the catalog to describe the catalog's observation-related properties. A data model for the source extraction process will also be required. For a heterogeneous catalog where each source may originate from a different Observation, we will probably want to define a standard simplified view of the Observation model that could be easily serialized as columns in the catalog.

6 Model

- The Observation model has three main parts: Observation Data, Characterization, and Provenance. Loosely speaking, the three parts are

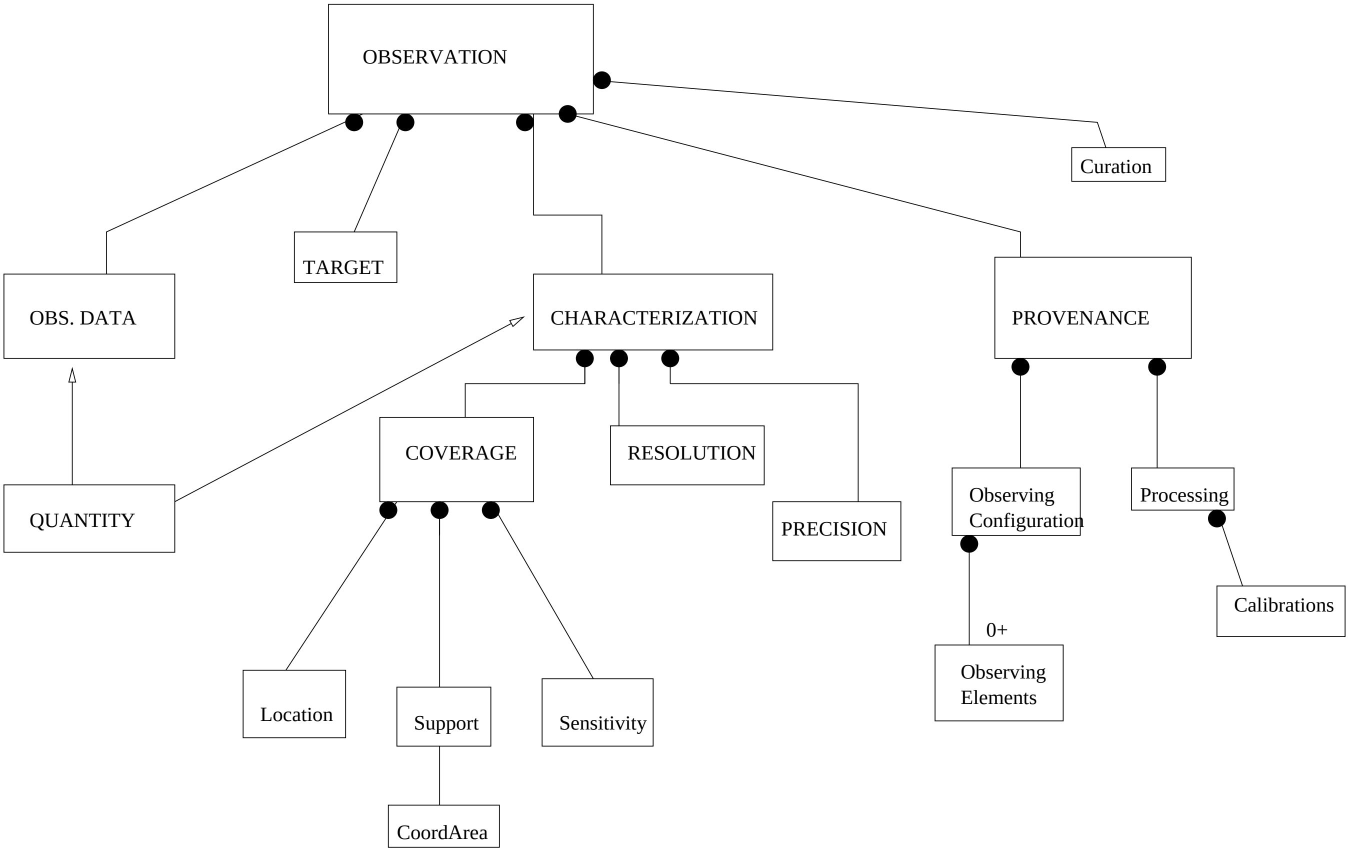

Figure 4: General model for Observation. Not all inheritances from Quantity are shown here.

metadata saying what the data is, metadata describing how to use the data in its current form, and metadata describing how the data was generated.

- Observation Data is a placeholder for the Quantity class (see the Quantity document, in work). It describes the axes and dimensions of the data.
- Characterization consists of Coverage, Resolution and Precision. The axes of Characterization are instances of Quantity. They define the different parameters constraining the data. These axes will be referred to as the Characterization space. Associated with this space are descriptions of the Coverage, Resolution and Sampling Precision.
  - Coverage describes the sensitivity of the observation along the Characterization axes.
  - The components of coverage are SensitivityFunction, Support and Location.

- Location consists of a set of coordinate values describing a representative location of the Observation Data in the Characterization space.
- Support consists of a set of CoordArea objects describing the extent of the (valid) data in parameter space.
- SensitivityFunction consists of numerical values indicating the variation in response as a function of the Characterization space coordinates.
- SensitivityFunction, Resolution and Sampling Precision have methods for returning a value for any coordinate in the Characterization space, as well as returning a characteristic value for the Observation.

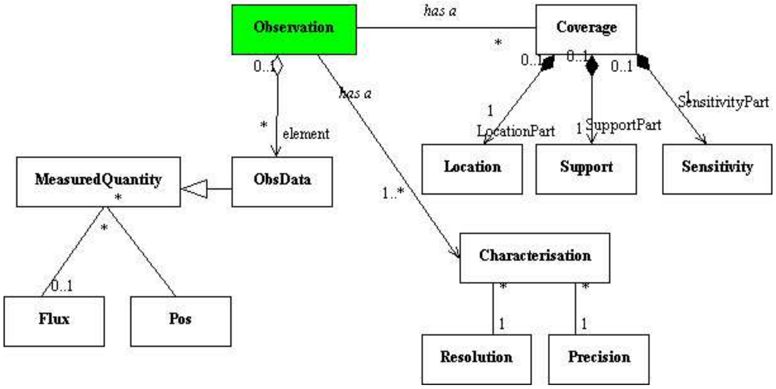

Figure 5: In this alternate view of the model Coverage has been removed from Characterization. We illustrate an instance of an Obs. Data as a simple Quantity object with a Flux array and Pos (RA,Dec) axes.

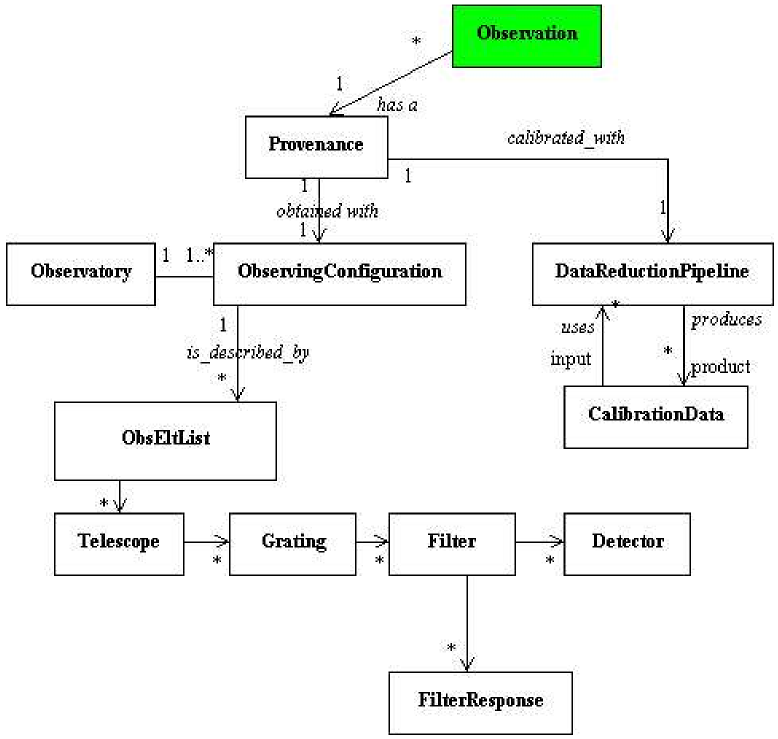

Figure 6: In this diagram we elaborate the Provenance object. Note that the specific ObsEltList shown here is an instance example and could be replaced with any arrangement of observing elements.

In Anita Richards' proposal, summarized in the adjoining figure, Provenance would also include Project (equivalent to Bonnarel/Louys Proposal), ObsCatalogue (equivalent to the Data Tree), and SourceSchedule objects. I have used a dashed line for ObsCatalogue, since I would like to suggest that as an archive model it should be outside the individual Observation model. *Discussion of how to incorporate these ideas is needed!*

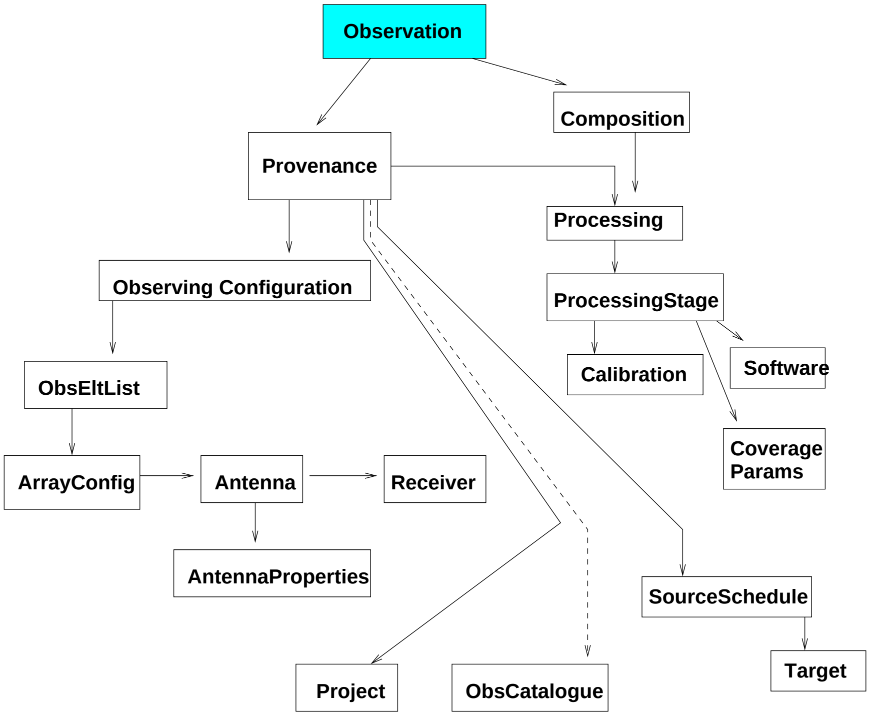

Figure 7: An alternative model for Provenance proposed by A. Richards for radio data.

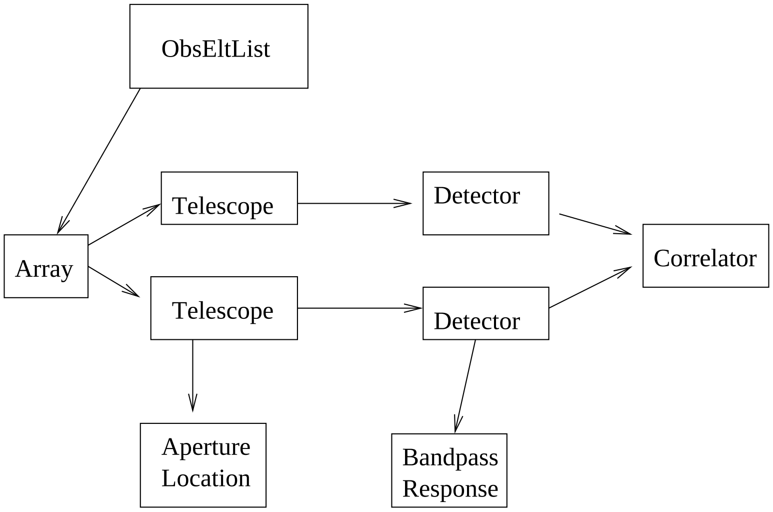

Figure 8: Another example of an ObsEltList, showing an interferometer and emphasizing that the list can have multiple branches.

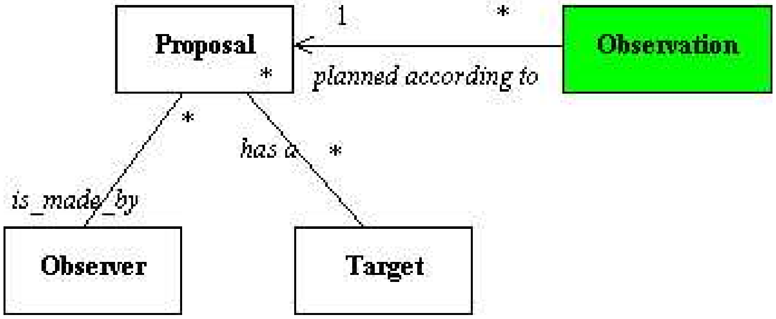

Figure 9: The model also describes the intent of the observation, considered as a Proposal. Two aspects of this intent are the Observer and the Target.

7 Packaging

Observational data belong to the real world, and each observation is accessed via files (in the most general sense - sometimes perhaps a URL constructed on the fly as a set of internet packets, but often via reference to real physical disk files). It is important that our model also should describe this organisation, which we refer to as the Packaging. The Packaging of an observation deals with the organisation of the dataset in files. Is the dataset a single file? A set of files ? are they contiguous ? Do they overlap ? What are the relative positions of the subfiles in the sub-frame.

Packaging also deals with the format used for your data. Although in the VO the trend is to describe metadata exclusively in XML, the data themselves will often be in other formats. The most common astronomical standard for observations is FITS n-dimensional images but it is also possible to find FITS binary and ASCII table extensions, ASCII text files, jpeg images, hcomp, GIF, PNG, or other proprietary image formats. Data in fully XML formats (either ”pure” XML or VOTables) is an additional possibility nowadays which should be possible to describe in our Packaging class.

Further, Packaging deals with the description of the coding algorithms used for the data: data compression, or redundancy reduction techniques. Discrete cosine transforms, H transforms, PMT and wavelets are the most common approaches. Quadtree and Huffmann coding, which generally come later in the coding process, need also to be described or referenced. We can model this by providing:

- the name of the Format, e.g. ’jpeg’, ’MRC’
- the name of the codec (compression algorithm), e.g. ’Jpeg-2000’, ’PMT/MR1’
- a reference url for documentation (eventually),
- a link url to download compression/decompression programs

Packaging therefore also deals with the size and compression rate of the data. This is important for predicting I/O and processing time as well as network transfer time and other resources.

Several of the items discussed here are not always independent. JPEG for example refers both to a coding algorithm and to a format. It has to be noted that each observation can support a very wide (infinite?) range of different Packagings and that in a given application an observation can support several different Packagings. This has consequences on the multiplicity of the links between the Characterization class and the Packaging one.