Data Model for Observation

DM Technical Meeting 4
ESO/Garching, 2004 Jan 26 and 31

Jonathan McDowell, CFA, Gerard Lemson, MPI, David Giaretta, RAL Francois Bonnarel CDS, Mireille Louys, CDS, Alberto Micol, ESO

Day 1

Agenda review:

1. - review the requirements for this data modelling effort:
2. - examine the domain DM from G.Lemson and P.Dowler
3. - position of Observation model with respect to Quantity data model.

1 Requirements

The goal is to provide a draft document describing the core structure of the Observation data model. Observation DM should deliver a model useful and understandable to other VO Working Groups: DAL,VOQL,... Metadata modelling can include different approaches:

- -a query oriented view, that will emphasise the properties and attributes of the data model, and require a dictionary, "ontology" framework.
- -a data processing point of view, science oriented that will define first the interface of data model objects

A spiral development is probably needed between both, using the interfaces provided by the Quantity DM.

2 Domain DM from G.Lemson and P.Dowler

This is exploring a way to go from concepts in astronomy to UML Classes, then XML descripton and/or Java classes at an API level. This matches the 3 phase process of an application development: Analysis-Design-Implementation and uses a top-down approach. The domain is shared in several packages that distinguish phenomenon, protocol, experiment,... it describes metadata about simulations: data + programs It is part of the GAVO initiative to model theoretical data.

GL tried the translation from the uml model down to a xml document

```
UML-XMI-(XSLT)-XML
```

Comments from DG: do not go beyond the scope of DM WG in trying to encompass the full VO architecture problem.

How we can build the model:

- -use cases should be collected and used as starting point for model elaboration, and in the validation process.
- -Different levels of granularity are needed: e.g: Registry uses a coarse grain description compared to what the user wants to retrieve.

3-the Observation DM:

The concepts to be modelled here have been listed in Strasbourg , Oct 2003 and reported in a IVOA ID (Internal Draft) by Jonathan McDowell:

```
http://hea-www.harvard.edu/~jcm/vo/docs/dmtm2.ps
```

This first version DM concentrates on the following

```
Observation components:
Curation
Coverage
Provenance
Observatory
```

Catalogs are not discussed yet but are a priority.

Reusability: Observation DM should provide hooks to the Quantity objects.

Modelling Quality: this is partly tackled in the Quantity DM, but should be detailed in Observation DM. What is quality? It's different from an error, closer to Accuracy (Quantity); How can we represent it? Options include

- confidence flags (GL/PD model)
- numbers that can be ordered
- things concerning the data (artefacts, cosmic rays, bad pixels, shutter on...)

Quality can be attached to any Quantity. It may have different levels (as confidence levels) e.g. solar regime data.(DG) Quality depends on the data type eg. FLUX

Use-cases for Observation

We proposed several kinds of dataset that Observation would represent: N-D image, Rosat Event list, data cubes : velocity cube, Mosaic images. We distinguish Observations from data products, like catalogs of sources, catalogs of lines. Observation will describe images, spectra, SED, data cubes..

Some discussion about the double meaning of the word mosaic: in surveys like 2MASS, it means an average or co-addition of images of an extended region, superimposed or only partially overlapping. It is also used to refer to detectors such as WFPC2, where it is a collection of several chips, or the tiling of those chips into a multi-chip image. We will reserve use of the word mosaic to imply the tiled or coadded image in which direct information about the original pixels has been lost; datasets with the individual chips still separated will be referred to as multi-detector images (MDI).

We must provide traceback information for mosaiced images. Must allow for a traceability value:yes or no. e.g. in the chandra detector: many different configurations of detector chips: any 6 out of 10 The model could be VO-compliant at diff levels:

- level 1:basic info is present
- level 2:intermediate info is present
- level 3:fully described.

mosaiced images stemming from diff instruments/telescopes.

How can quantity be used for a MDI image?

```
outer layer: give me the pixel values at position alpha, delta
inside: chip nb, X, Y
Quantity allows for alternative representations.
OBS= list of qty objects (one per chip), with PSF attached
methods required:
 interpolate flux at position P
 list me the data available at pos P,e.g the overlapping raw images.
 do the interpolation with my parameters.
metadata needed to do that:
 coords system on each chip
 list segments of the detectors
 values at a particular array element
 accuracy of the astrometric solution.
Eventlist
 Energy Qty
 Photon Datatype
 list of photons = 1QTY with energy position.
methods: give me all the info about each photon.
```

Coverage:

Circulated Alberto Micol's document Data Characterisation (DM list,nov 6, 2003). Discussed nomenclature. (1) support :where we got data ,region that are covered by the mission. (2) sensitivity: response of the flux measured in these regions.

The problems: spectral leaks, variable PSF along field of view, etc. Sensitivity as a function of : energy, position, pointing path (trajectory),time. Need to code the aspect solution (pointing direction versus time). What is the relative exposure ?

	TIME	POS	ENERGY	Observable	"X"
				flux	simulation
support	good time intervals	FoV	bandpass	saturation level limiting flux pile up	defining range
sensitivity	sensitivity	quantum efficiency vignetting	transmission effective area		
resolution	perfect?	PSF	line spreadfunct.	?	
pixelization	integration time of CCD frame size sample interval	pixel size	channel size		model grid
mapping detector-to- world coords	clock rate (reference time)	WCS	gain	count->flux linearity	grid
Filling factor	dead time	pixel factor	?	?	?

Observatory

Provide a general case for many kinds of instruments used in the different regimes. Cases: optical , radio, telescopes, spectrometer, X detectors. full photon(EM wave) path to be modelled.

Coords for telescope: nominal, effective , ground based vs orbital coords. background

Source and background images; Define background as a residue from a selection process: e.g source extraction. in contrast, the sky images,flat field images, superflats are additional information to the fully calibrated image. they may share common metadata with the observation : ex pointing position, dimensions, size, even astrometry (?) but not necessarily. however such calibrations are to be delivered to users if they want to re-calibrate the data by themselves. (NB also a requirement from the science WG)

Calibration provides a link to calibration/ancillary data and calibration programs description as well. describe the calibration process, distribute the metadata to evaluate the level of calibration and sensitivity/error of the calibration process or full acquisition +calib if not possible. flux calibration : required is the link to applied standards.

provenance cannot be extensively tackled at present. Methods needed to add provenance elements in the descriptions of observation / programs ,etc... as a continuous process. (template)

Day 2

1- Scope :

Postponed for next iteration: (1) - data Quality modelling, but provenance which is related has been studied extensively. (2) - Serialisation (3) - catalogs, in close links to sources, objects, scientific content.

Thread, Philosophy of the document: The DM should reflect how the metadata are queried in the retrieval process of astronomical data, on one hand, and express the user's approach for data interpretation on the other hand. Science cases are definitely useful in this perspective.

Gerard suggested the use of views in the modelling effort. eg. the metadata tree is a view of the data collection opposed to the 9-axis view of a data cube.

2-Review of the extended DM Observation after Monday discussions.

0.0.1 UCD and data models:

GL: define enough new UCDs to map on the data model elements. ML view: UCDs as they are now can be used as semantic tags at the attribute level. Specifying UCDs at the level of the OBS object is simply not possible and too controversial.

Noise: should we discuss this in the DM? difficult to define in general. Where do we make the cut: some signal for a user will be noise for the other. Depends on the aim of the user in the analysis of data: e.g, CMB, sources extraction, deconvolution, etc... The characteristics of the instrumental noise should be there, attached to the Provenance objects. The place holder is still to define and to validate. For example, the Read-out noise could be included as parts/attributes of Detector,as well as Gain, Quantum efficiency.

0.0.2 Coverage:

We have tried to insert the coverage in the UML diagram . A general Coverage object is there that matches the principal properties listed in the characterisation table (cf Dm meeting session 1, jan 27)

Objects corresponding to the rows in the columns of the characterization table are hooked below the Coverage object. Location, Support, Sensitivity.

AM comments: location/support may be dependent/related to sensitivity. weight maps, calibration maps, etc... ex determining the position requires sensitivity knowledge. sensitivity may depend from the location/support, e.g MDI images, (multi-detector images, non uniform transfer function, etc... Characterisation of the data is an Object as such. It summarises many aspects/properties listed in the characterisation table. Resolution, Pixelation, etc are connected to it at this level and not at a more internal level just to be easily extracted from the Observation Object. (query orientation?)

The alternative would be to store it in an Accuracy object at the level of the data container (Quantity),or to store it in a Coords object as suggested by the STC strategy. For the sake of simplicity and access performance, Characterisation properties needs to be attached at the top level of Observation.

For complex observations, multiple files/Multiple detector images: A Characterisation object can be attached to each element (spectrum, plane, N-d cube, ...)

Alberto: Has this something to do with Packaging ?

0.0.3 Instruments

Include a class like Aperture. A way to homogenise the different regimes ? an Aperture inherits from the Observatory class (location information)

0.0.4 Calibration:

Use case:

I have retrieved a spectrum and want to know what are the processing already applied to the data. Have my retrieved data already be de-reddened or not? GL argued to use a work-flow approach to express the dependency of the different steps. This kind of thread could map the complete provenance and express the different steps of calibration process or data analysis process. JCM argued that this is too inconvenient and one wants to focus on the nature of what the data is now, not describe that by what has been done to it.

How do we hook the calibration parameters and calibration output data? This is good to express the diff tasks applied to the raw data (reduction pipeline) or data analysis pipeline as well. But to answer the former question, specified steps as "de-reddening" or result as "de-reddened data" should be defined. This mapping between the process and the physical interpretation is the harder job, and can be done only for very well specified tasks, already designed in reduction pipelines. Start to think about "corrections" for each step in the work flow.

How far should the VO explicitly describe the data ? The most general procedures should stay simple. The full reduction data and procedures description need a big amount of sophistication that will add an overhead to the regular user. How do we match /tune the requirements between the query view and the physical interpretation view, somewhat beyond the data analysis.

0.0.5 Target:

Related to observation. could be a link to a Target Object. the subject of the protocol in GL domain model.

0.0.6 Background

validation of complex observation with a reduced data set expressing the target , plus related observations, taken acoording to the same provenance and used as complementary data,for later analysis, quality assumption, etc. e.g. a spectrum , plus a background spectrum taken in a close/near-by region. JCM emphasized the loose use of the word 'background' by observers to mean 'observation of region B used as a model for the background at region A'. Spectral archives often retain source-background pairs.

This pleads for an Obs data seen as a bunch of files including calibration shots of the same Field of view.

STC discussion

We studied the STC UML document and browsed the diff levels of description. We did not come up with a full mapping of Quantity/STC objects. The Axis object in Quantity encompasses rather well the STC/Coords family objects but more work is needed here.

Use cases

Proposal and Observer: are part of the Experiment in GL model. Should be linked to the Observation class.

Call for use cases : - query use cases - scientific use-cases ,which means here , with a physical interperation scenario for the retrieved data.

- 1) find quasars with Xray jets select objects with a defined energy band ,2-10 keV and Depth ¿ 1 nJy.
- 2) where in the same region may we have more than 4 quasars. – requires source/objects models attached to images ingredients : POS + bandpass + some known catalog results.

- 3) Quasar SED composition - find all spectra at a given position, correct each one using a host galaxy model using the aperture size of each spectrum.

- 4) make an Sextractor use case cf AVO-demo Which parameters do we have to provide to the ACE interface. Should be queried in the data set against the DM. What we need is a constructor method for an Sextractor parameter file. - input could be : -simple parameters (param=value) list -parameter + weight maps cf Sextractor config + weight maps. Ingredients : POS, aperture,time, restframe

Notions to be there:

reduction level, zeropoint, limiting flux magnitude error: in data characterisation ? with a hook to Qty:Accuracy

N.B: needed is a full example of an Observation Object with the Qty data container. - images, - spectra, - event list : aspect solution , Qty, etc...