Data Models for the VO

Jonathan McDowell

Smithsonian Astrophysical Observatory, 60 Garden St, Cambridge, MA 02138, USA

1 What Is a Data Model?

The key to making the Virtual Observatory work is the definition and adoption of interoperability standards. One form of such standardization is to agree on exactly what we mean by the data objects we all deal with - images, spectra, coordinate systems, etc. This standardization and abstraction process is called data modelling. We have been using data modelling techniques for several years to develop the CIAO analysis system for the Chandra X-ray Observatory [1].

A data model [2] is a recipe to describe 'how is my data different from (the same as) your data?'. By different, I mean in terms of abstract information content rather than specific byte format. Suppose you have an simple 2D image of part of the sky, you can store it as a FITS file or a GIF image and you've got the same information – until you add a coordinate system to the FITS file or a color table to the GIF. Now suppose the image was made by mosaicing four chips and you have a FITS image with one extension per chip. There's new information – your display program may show the same picture but you have retained the information of which part of the sky is observed with which chip. By elaborating a data model that describes astronomical images we ask: what questions can I ask about an astronomical image? In this case, for instance, how close is this star to the edge of a chip?

The data model describes the information content, and the metadata protocols that Ray Plante discusses in the following paper describe the way that content should be formulated and tagged – the boundary between the two is a bit blurred. The data model may also describe the access functions ('methods') for the data.

For many VO uses, catalog federation is all you need to do and that can make do with a fairly simple data model – although the issue of sky coverage is tricky, and Arnold Rots addresses that in his paper. For VO applications that work with image and spectral data directly, data fusion work, a good data model is much more critical.

In the context of the VO, the VO consortia will use the data model to design the metadata, making sure the most general image can be represented. A data provider will use the image data model to map their data to the standard VO representation, and tell us which questions their images can and can't answer.

Data models also allow you to compare disparate types of object. All astronomical data has some commonalities (the need for keywords, coordinates). A

ESO Symposia: Toward an International Virtual Observatory, pp. 101–105, 2004.

data model for images or for spectra can be considered as a special case of an astronomy object; This is just another way of saying that we shouldn't implement coordinate systems for spectra and images in two totally different ways.

2 Not So Easy: A Real World Example

Let's get specific and look at the example of four images I got out of four different archives (Fig. 1, Table 1) – a ground based telescope, HST, Chandra and ISO. Not one of these is a simple FITS image, so when you get them back from the VO if they're in their present form it's a lot of work to combine them – **no one software tool will operate correctly on any two of these images!**

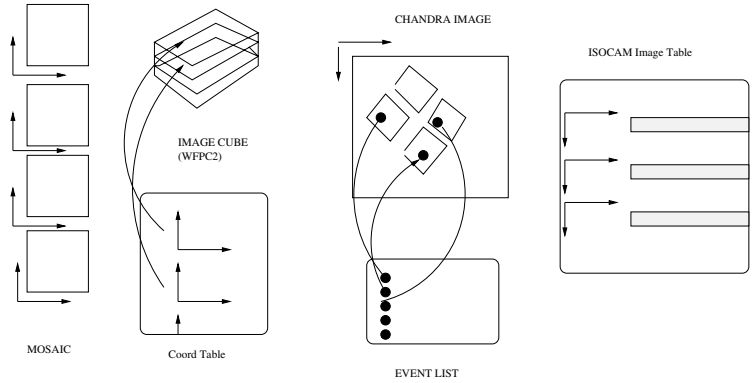

Fig. 1. Different FITS implementations of a mosaic image. The arrowed axes represent coordinate system metadata

Table 1. Structure of mosaic images in different archives

	FLWO Mosaic	HST WFPC2	Chandra ACIS	ISOCAM
HDU Structure	4 images	1 3-D image	Event pixel table	Table of images
Coordinates	WCS keys	Special table	WCS keys	WCS columns in table
Bandpass keywords	FILTER	PHOTPLAM	DSVALn	WAVELENG
Duration	EXPTIME	EXPTIME	LIVETIME	DATE-OBS, DATE-END

It's not just a matter of different keywords – it's easy to map those. It's a matter of different approaches to encoding equivalent information content. For instance, X-ray data doesn't have just a start and stop time, it has a table of multiple start and stop times, making it harder to answer the question 'did this star flare during my observation'. WFPC data has standard WCS coordinates on the 3-dimensional image cube, but they're misleading about all but the top plane and the real WCS values are stored in another table, while the logically equivalent ground based and ISO mosaic images take two further different approaches to encoding the same information.

It's not enough to map each of these examples directly to a VOTable [3] in their current structure. That won't capture the fact that they contain the same kind of information – we have to define how to give a uniform structure to these data. So for the VO we need a standard for mosaic images, a standard for coordinate systems, perhaps a standard for timing information.

Some things are more important than others – it'll be much more common to need the wavelength than the observing location. We can model the main things first and develop more standards as time goes on: the VO will become aware of the answer to more and more questions.

The data model will let the archive provider figure out what they have in the common language of the VO: 'mosaic image with one coordinate system per image'. The metadata standards will tell them how to represent such a thing in, e.g. a VOTable. VOTable has sets of nested tables, like most formats it has keywords, and it has a simple object for defining celestial coordinate frames. There are things FITS has that it doesn't yet, like coordinate transforms and image axes, and there's no structure to the header. In FITS, you have a set of images or tables each with a set of keywords. There's more implicit structure created by defining objects with groups of keywords. We will need a more sophisticated and explicit structure to describe the data we will analyze with the VO.

3 Modelling Data and Metadata in the VO

The NVO group is discussing a VO data model in which a dataset (for instance as represented by a table in a VOTable) will contain a set of columns and/or images with a set of metadata descriptors (Fig. 2; the model is presented in more detail in a discussion document available in the NVO document repository, currently at `http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/vodm003.ps`). Where in FITS we have simple keywords as the building block for the header, in the VO each of the metadata descriptors is a whole object – this may be as simple as a keyword or it may be quite complicated, and as an object its definition may be extended as time goes on. Also, descriptors may be attached not just to the dataset as a whole, but to individual columns or images or even to other descriptors. I call out a special case of coordinate descriptors here attached to the image, since they are so important. In the figure, 'hypercube data' refers to the N-dimensional image data or the table column data, as appropriate.

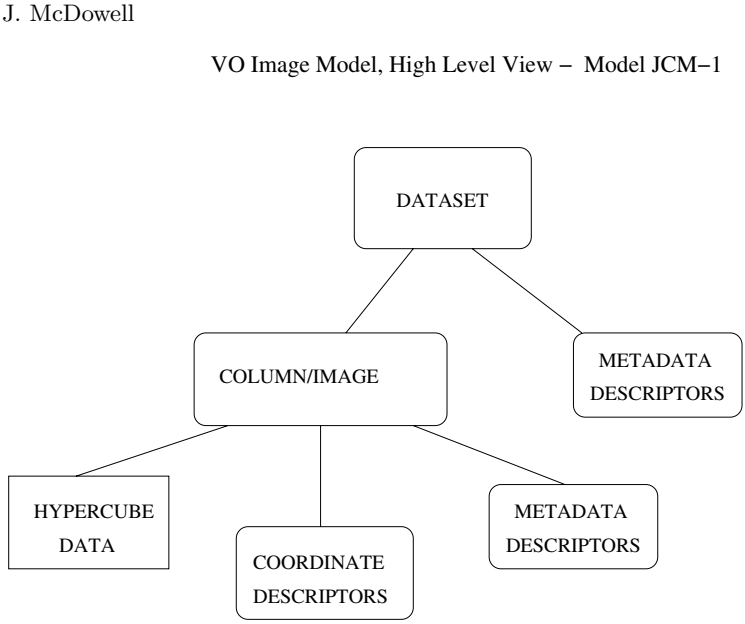

Fig. 2. Proposed overall image/table model

In Fig. 3 I've illustrated a set of metadata objects we have identified as worth modelling. Let me focus on three related objects to conclude this presentation.

- Data Quality: this comes up in all data analysis systems. We should define a common approach to describing bad pixel masks, quality flags, observing interval interruptions, and other kinds of lacunarity. Quality isn't just on/off as the definition of 'bad' may depend on the science; exposure depth folds in here too.
- Data Subspace is a concept we introduced in the Chandra data analysis system CIAO, to unify the answer to the question 'what range of time, energy, sky was this dataset taken from?'
- Data Fidelity is a new idea that I'm proposing here, it's slightly different: what level of correction has been applied to the data, where on the slider bar from raw, instrument-space data to unreliable, heavily modelled calibrated data do you lie? In the VO we'll eventually need to be able to specify this at some level. (In the discussion, Andy Lawrence pointed out that 'fidelity' is used with a different meaning in other astronomical contexts, and so a better name is solicited).

The way forward is to talk about these issues and compare the different approaches used by different archives and data analysis systems.

This work is the result of extensive discussions with the CfA VO team, the NVO collaboration members, and the CDS/Strasbourg VO team.

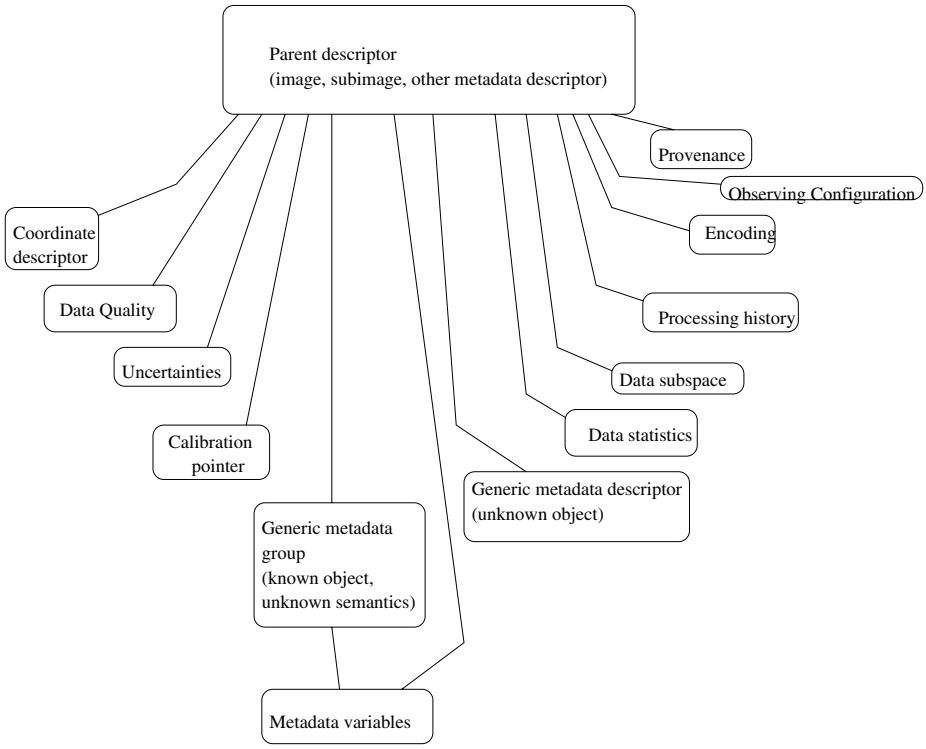

Fig. 3. Objects we will need to model

References

1. J. McDowell: Proc. SPIE 4477, 234 (2001)
2. A. Farris, ADASS 2, 145 (1993)
3. R. Williams et al., `http://cdsweb.u-strasbg.fr/doc/VOTable` (2002)