Spectral Data Models for the VO

Jonathan C. McDowell (CfA),

INTRODUCTION

A significant fraction of the public data currently available to the astronomical community is in the form of spectra. Unfortunately, different projects use very different formats and representations to publish such spectra. The Virtual Observatory will need to specify a uniform way for data providers to describe their archived spectra to users.

This study attempts to isolate the metadata needed for representing spectra to the Virtual Observatory, and proposes ways to structure that metadata. The ultimate representation of spectra should be a special one-dimensional case of an n-dimensional image object, but an interim spectral model ensures that we capture spectral-specific metadata and can later check that our n-dimensional model is an adequate generalization.

Our model separates metadata needed by applications using the idealized, generalized spectrum (pixel values, coordinates, errors, units, resolution) from metadata describing the idealized observation (sky region, observation date) and from metadata which is needed by specialized applications which deal with particular observational strategies (e.g. details of spectral extraction from a 2-dimensional imager).

WHAT IS A SPECTRUM?

- We will mean by a spectrum the value of an observable (usually intensity in some sense of radiation) as a function of a (photon) spectral coordinate (wavelength, frequency, energy, etc.), corrected or not for various instrumental effects.
- Distinguish between a spectrum in the theoretical sense, the energy output versus e.g. frequency $F(\nu)$, and a **spectral dataset** in the observer's sense of 'taking a spectrum', which maps such a spectrum onto an instrument in often complicated ways (echelle spectra, long slit spectra on an imaging detector, etc.).
- Spectral datasets often have the unpleasant property that three axes (celestial coordinates and the spectral coordinate) have been projected onto two instrument coordinates, introducing degeneracy in the data. In this document I will describe spectra (the idealized $F(\nu)$) rather than spectral datasets, but keeping in mind the complications introduced by those datasets - for instance, long slit spectra force us to immediately consider arrays of spectra as a function of a single positional coordinate.
- The 1-D spectrum as discussed above is clearly a special case of a 1-D histogram, and our final VO scheme should unify common metdata with other 1-D histograms (e.g. lightcurves) and with n-dimensional generalizations such as the 2-D image. This case study will be used to ensure that the n-D observation model can encompass everything we need to represent a spectrum.

OTHER KINDS OF SPECTRUM

1. Other observables as a function of wavelength: percentage polarization, extinction coefficient. These can use the present model.

2. Arrays of spectra such as spectral-spatial data cubes. We don't consider these here, but they are a simple extension if we model spatial images compatibly.

3. Spectral coordinates for particles other than photons: massless (gravitational waves) or massive (electron energy dist. in radio jet, cosmic ray spectrum).

4. Spectral coordinates not a particle property: power spectra of source variability or CMB anisotropies, Fourier transforms in general. Needs a slightly different model.

EXISTING AND EMERGING STANDARDS

The FITS WCS community is in the late stages of specifying standards to describe the mapping of pixels to a wavelength, velocity or frequency axis. However, there is no general standard, in FITS or elsewhere, for the organization of the pixels themselves. Doug Tody has recently carried out a survey of spectral archives (www.ivoa.net/forum/dal) for the VO which revealed a heterogeneous collection of formats, many in ASCII tables, FITS tables, or FITS images. This is in contrast to the situation with simple sky images which, despite problems with how to represent mosaics, are mostly in some variation of FITS image extensions.

OBSERVABLES

A crucial task for the VO is to standardize how data providers describe the observable. What do the pixel values represent? At the moment, if you are lucky there is a BUNIT keyword in a FITS image to at least tell you the unit, but that's not really sufficient. The VO will use tags such as Uniform Content Descriptors (UCD2, discussed elsewhere at this meeting) to unambiguously characterize the physical concept being measured. Our spectral data model must define a standard place to store this metadata.

Observable	Typical unit
Energy flux Density vs λ	erg cm^{-2} s^{-1} $\mathring{A}^{-1}$
Energy flux Density vs ν	Jy
Energy flux Density vs $\log\nu$ (for SED)	Jy Hz
Photon flux density vs Energy	photon cm^{-2} s^{-1} keV^{-1}
Luminosity (at source)	erg s^{-1} $\mathring{A}^{-1}$
Luminosity per decade	$L_\odot$
Radiation energy density	erg cm^{-3} Hz^{-1}
Flux per solid angle (e.g. at source surface)	erg cm^{-2} s^{-1} $\mathring{A}^{-1}$ sr^{-1}
Antenna temperature	K
Brightness temperature	K
Magnitude in given band	mag
AB magnitude	mag
Surface brightness flux density	Jy / arcsec^2
Flux per resolution element	Jy / beam
Surface brightness mag.	mag / arcsec^2
Instrumental reading	ADU, count
Ratio of two spectra	Dimensionless

Table 1: An incomplete list of spectral observables

SPECTRAL PARAMETERS

The spectral survey confirms that existing public data use the full range of possible parameters used to label the electromagnetic spectrum:

- Frequency, wavelength, energy, wavenumber
- Base 10 log of these quantities

- Various kinds of velocity

A PARTIAL MODEL

The model displayed here is an elaboration of one circulated to the VO community in May 2003. The boxes indicate how we might structure the metadata for spectra, but the model is general in the sense that by adding additional axes to the data container it could be applied essentially without change to N-dimensional images. The details of the model will change as other models such as Quantity are fleshed out.

There are three main parts of the model: the dataset description, the data container description and the observation coverage description.

- The first diagram shows the complete dataset, which contains curation and coverage objects as well as several Data Container objects. The dataset will have at least one Data Container for the main data, and may have additional ones for a background spectrum, an exposure array, and a sensitivity array.

- The Data Container (second diagram) has a Data Storage object containing Value, Error, Quality and Resolution sub-objects.

  Our abstraction is that the data consists of an ordered array of values (accessed by the Index object) which may be coupled to one or more PixelMap objects locating each value in a coordinate system (see the poster by Lowe et al. for more details). In the spectral case, the PixelMap would provide a bijection between pixel number and the spectral coordinate. A simple case of such a map is a set of regularly spaced, contiguous wavelength bins. However, our abstraction also supports irregular or sparse arrays.

  One may in general obtain value, error, quality and resolution numbers for each pixel, although in many cases things like the resolution may be constant for all pixels; the four separate objects, accessed using the Index, hide this implementation detail.

- The Coverage (third diagram) is a simplified summary of the Space

Time Metadata of Rots et al. (hea-www.harvard.edu/∼rots/nvometa) and encapsulates the spatial and temporal region from which the spectrum was extracted.

DESIGN ISSUES

- The observable is declared with the UCD attribute of the Data Storage object. We need to elaborate this to fully model a Photometric System object.

- The resolution is grouped within the Data Container together with values and errors, emphasizing its essential role in the abstraction. The resolution object should be a line spread function at each pixel.

- In contrast, the sensitivity (counts to flux), exposure and background are treated as separate data containers for two reasons: firstly, their effects are considered to be calibrated out, and accounted for in the error object; and secondly, they often have their own error, quality and resolution information different from the main data - although we should require them to have compatible pixel maps in some (to be made precise) sense. Alternate choices would be to include all these arrays in a single Data Storage object, or at the other extreme to consider them as separate but associated Dataset objects and replicate all the observation information.

  The sensitivity and exposure require particular care when we extend the model to a 3D energy-position cube, where practical implementations are likely to express things separably as, e.g., an on-axis energy sensitivity and a spatial sensitivity map.

- UCDs will help us describe what corrections have been made to the data, but our model does not yet explicitly have a way of specifying that a spectrum is in the rest frame and corrected for Milky Way but not intergalactic absorption, or corrected for detector QE but not telescope vignetting. This should probably be part of the observation description, but one might argue it belongs in the data description instead.

LINELISTS

A common form of archival data containing spectral information is the line list, a catalog of observed lines and their properties such as equivalent width, FWHM, integral flux, central wavelength, and identification. Such a list implies, and can be used to create, a spectrum in the same way that a source catalog can be used to reconstitute an image. We choose to model this with the idea that line lists and source catalogs are objects that are not themselves spectra and images, but which have methods which map them to spectra and images. In other words, we will build a line list model which is separate from the spectrum model.

The essential feature distinguishing the entries of a line list from the pixels of a spectrum is that each entry is thought of as representing a distinct physical process in the source which could at least potentially be identified with a transition of some kind (C IV λ1549, and so on). Secondarily, the fluxes reflect integral properties over a finite range of the spectrum rather than a measure of the monochromatic flux density at a single resolution element. (It is possible that some X-ray spectra fits best represented as integral fluxes might share the line list model). To map a LineList object to a Spectrum object, one needs to assume a line profile (to go from integral to differential space) and discard the identification information (in our model, the Spectrum object does not have identified features; for display applications one might want both a Spectrum and an associated LineList).