Compliant Data Models

- First step: a text white paper defining concepts
- Second step: UML Class diagram with:
  - Versioning
  - UCDs (or successor)
  - Full text description explaining attributes

  To avoid huge diagrams, we recommend a set of nested diagrams with only half a dozen boxes per page.
- Third step: Provide a reference XSD schema and XML instance examples. Recommended XSD created by a tool to ensure validity. The XSD provides a reference representation and the examples clarify the intent.
- IVOA DM working group to adopt models; should have 2 groups exchanged data using the model before final adoption
- Classes do not have to include associated methods (mainly just define attributes); if methods prescribed, must include relevant UML activity diagrams and text description of arguments and algorithm.

Listed a bunch of objects one wants to model and picked out a few with attention to those needed by DAL WG. Models listed: - Bandpass resolution lim flux time of obs Roadmap for WG: - Continue elaborating and converging 'big picture' data models (IDHA, CVO, Astrogrid, etc), but don't expect completion soon. - Work on targeted smaller models which can be completed in a short time. Assign leaders for each such model. - Work will be done on dm mailing list, probably supplemented by monthly telecon - Each package to use subject headers on email e.g. [SPECTRA] - Promote reusability of model components (eg Quantity, STC) dave g - interfaces more important than classes? but this is for dal wg? jcm NVO wants spectral model this summer transform including wcs and units Priya 4:30 heffers [] QUANTITY Ray (Norm Gray)... SPECTRA JCM Pierre, ML RESOLUT Pat, Doug TIME_OBS Pat, Doug TRANSFORM Dave Berry? later INTERFER Peter Lamb SIMULS Gerard... OBSERVATIONS IDHA+CVO+?

Some comments on UCDs

- The same UCD can apply to two different columns in a table
- UCDs are not fully precise (same UCD refers to two different concepts)
- UCDs are a flat tree, but the form of their names does imply a hierarchy (FOO_BOO_GOO)
- UCDs are far from complete: e.g. attempt to map X-ray data file header keywords (Temporal Observation Coverage absolute and relative uncertainties, start and stop times in mission seconds, disperser/grating name, subdetector name, target source position as opposed to telescope aimpoint, spacecraft roll angle, different kinds of instrumental coordinate, instrument energy gain scale, event grade pattern, instrument bad pixel mask, coordinate system reference pixel, etc, etc).
- General agreement that 'atoms' needed for UCDs such as 'ERROR' (need to specify error of what other UCD)
- Astronomical concepts not well mapped by discrete tags - many concepts have continuous (real-valued) components. Example:
  - PHOT_FLUX_RADIO_1.4G

  but what if you have 1.5GHz instead? Prefer
  - two UCDs PHOT_FLUX and FREQ_GHZ (=1.4), and
  - some way of declaring their relationship.

PCDs

To drive discussion, I proposed an alternate vocabulary for applications which require precise interoperability of astronomical data.

- PCDs are more flexible, but more complicated than UCDs
- PCDs take other PCDs as functional parameters
- PCDs are structured objects, not a flat tree.
- PCDs are appropriate for labelling an object model for astronomy
- Some questions:
  - Are there requirements for precise semantics which the UCDs do not match?
  - Is it better to extend the UCDs to meet those requirements, at the risk of making them meet their existing requirements less well?
  - Or, is it better to use a different set of tags to meet a different requirement, at the risk of some redundancy?

Hierarchies

- There are many possible hierarchies for the astronomy problem domain
- The UCDs provide one such hierarchy
- My PCD document

  http://www.ivoa.net/forum/ucd/0016.htm

  proposed another such hierarchy
- Are all hierarchies equally good? No!

  Each astronomer's personal mental model is different, but there are many commonalities. It is easier to search the tree if it is organized in a way that makes sense to most astronomers.
- I believe this means the tree should be organized in an object-oriented way.

PCD/UCD comparison

UCD	PCD	Comment
ID_CATALOG	SRC_ID_CAT	ID of a source
ID_PLATE	OBS_PLATE	Property of an observation: which plate?
EXTENSION_RA	SRC_REG_DIA_RA	Property of a source: special case of a source diameter, which is special case of a source region
VELOC_CMB	OBJ_VEL(FRAME=CMB)_R	Velocity of what? Radial component? In what frame? (infinite number of possible frames)
POS_HC_X	POS(ORIGIN=HELIO,FRAME=ECL)_X	Which frame? What origin?
PHOT_FLUX_RADIO_15G	PHOT_FLUX(FREQ=15.0GHz)	Freq is a variable
PHOT_PHG_B	PHOT_MAG(SYS=PG,BAND=B)	Don't enumerate all combinations
PHOT_PHG_B-R	PHOT_COL(SYS=PG,BANDS=B-R)	Color and mag are the different concepts; which system and filters are qualifiers but not different concepts

- Possible alternate syntax (R. Plante): Use XML XPATH
- PHOT/COL[SYS='PG',BANDS='B-R']
- OBJ/VEL[FRAME='CMB']/X

Refactoring the concept space

- I propose a new hierarchy in which the top level domains are organized as:
  - Information about the universe
  - Information about observations of the universe
  - Information about datasets and their analysis (and about astronomers and literature and so on)

Primary objects: the universe

- PHYS - Physical data which is true generally, not specific to an astronomical location or object.
- POP - Aggregate properties of a population or class of astronomical objects.
- SAMPLE - Properties of a sample of astronomical objects
- OBJ - The main things we study are (more or less discrete) astronomical objects in space
- MED - We also study the physical medium making up these objects: consisting of matter (solid, liquid, gas, plasma), radiation, and fields.
- LOS - Sometimes we are concerned with the properties of a line of sight long which objects may line (e.g. total column density)

Primary objects: observations

- SRC - a subset of a dataset can be deemed to be a source. Sources may be identified with actual astronomical objects, but we distinguish between the direct measurements (flux, counts, angular size) of a source and the physical properties (luminosity, mass, diameter) of the object they represent. Is this too subtle?
- BKG - some sources (regions of the sky in a dataset) are called 'background' and are not identified with objects.
- OBS - metadata describing the observing instrument, the configuration, and the coverage (field of view, times, bandpass).
- SURVEY - a group of observations with some common properties.
- EVENT - a response by a photon counting instrument, putatively corresponding to a detected photon.

Primary objects: data analysis

- DATA - things about the datasets and files
- REFER - bibliographic info

(one could also imagine sociological metadata such as JOB_RUMOUR)

Primary objects: the universe

- PHYS - Physical data which is true generally, not specific to an astronomical location or object.
- POP - Aggregate properties of a population or class of astronomical objects.
- SAMPLE - Properties of a sample of astronomical objects
- OBJ - The main things we study are (more or less discrete) astronomical objects in space
- MED - We also study the physical medium making up these objects: consisting of matter (solid, liquid, gas, plasma), radiation, and fields.
- LOS - Sometimes we are concerned with the properties of a line of sight long which objects may line (e.g. total column density)

Primary objects: observations

- SRC - a subset of a dataset can be deemed to be a source. Sources may be identified with actual astronomical objects, but we distinguish between the direct measurements (flux, counts, angular size) of a source and the physical properties (luminosity, mass, diameter) of the object they represent. Is this too subtle?
- BKG - some sources (regions of the sky in a dataset) are called 'background' and are not identified with objects.
- OBS - metadata describing the observing instrument, the configuration, and the coverage (field of view, times, bandpass).
- SURVEY - a group of observations with some common properties.
- EVENT - a response by a photon counting instrument, putatively corresponding to a detected photon.

Primary objects: data analysis

- DATA - things about the datasets and files
- REFER - bibliographic info

(one could also imagine sociological metadata such as JOB_RUMOUR)

Data Models in the IVOA

- To exchange complex data we will probably use XML structures in files and (Java/C++?) object classes in software.
- A data model is an abstract representation of the class that isn't laden with implementation (XML or Java) assumptions.
- It's only useful if you actually implement it. How do we get there?
- We had a small meeting in the other Cambridge last year which came up with a proposed approach.

Compliant Data Models

- First step: a text white paper defining concepts
- Second step: UML (or equivalent) based model (diagram)
- Third step: Interchangeable format model (XMI?) satistfying IVOA DM standards (see next slide) - a 'Compliant Data Model'.
- Fourth optional step: Provide a reference XSD schema and XML instance examples. IVOA/Interop can adopt model as a **recommended VO model** for the given concept if it fits in well with other such models and is suitably extensible.
- Even if you don't do the 4th step, sites can agree bilaterally to interchange data using a specialized compliant model.

Conforming Data Models

Must have:

- Unique URI and name
- Version of model
- Descriptive text
- Curation metadata
- URL for white paper
- Class descriptions (with name, attributes, relationships and at least some methods).

Data Model for spectra

Existing resources:

- Greisen et al WCS paper: http://www.aoc.nrao.edu/~egreisen/scs.ps
- Bushko memorandum: http://www.ivao.net/forum/dm/0034.htm
- Bandpass white paper: http://hea-www.harvard.edu/~jcm/vo/band.pdf

What is a spectrum?

The platonic spectrum:

- 1-dimensional curve, or a set of samples of that curve;
- radiation intensity (flux density, luminosity, number of photons)
- as a function of photon spectral parameter (wavelength, energy, frequency);

- We want to include spectral energy distributions (SEDs), which are typical combinations of individual photometry points and continuous spectral arrays. There are three possible approaches:
  - consider photometry and spectra to be different types of object; an SED has N photometry points and M spectra with a total of m samples in them.
  - consider each point in a spectrum to have the same metadata as a photometry point, and ignore the fact that the samples in the spectrum are more related to each other than they are to the photometry values. Then an SED has N+m spectral values, with no further substructure.
  - consider photometry points to be a special case of a spectrum with m=1 sample each. Then an SED has (N+M) spectra with a total of (N+m) samples, N of the spectra have only 1 sample.

  I propose the last approach since it lets us share metadata for a given spectrum (e.g. slit size) and minimizes the different kinds of software object.

What is not a spectrum?

For the purposes of this white paper the following cases do *NOT* fall under the definition of `spectrum':

- A 2D spectral dataset (e.g. long slit, echelle) from which a spectrum may be extracted
- Other kinds of observable vs. wavelength: e.g.
    - percentage polarization or
    - extinction coefficient
- spectral parameter is energy, frequency or wavelength of particles other than photons, e.g.
    - gravitational wave spectrum,
    - electron energy distribution,
    - cosmic ray spectrum
- spectral parameter is frequency of something other than a particle,
    - power spectrum of variablity
    - power spectrum of CMB density fluctuations
    - tenure probability versus publication frequency

Basic properties

- The spectrum F(p) gives the **value** of an **observable** F as a function of a **spectral parameter** p. The attributes of the value will be those of a generic **quantity** and will include at least a unit, uncertainties, and support for upper limits.

- The parameter p can be any parameter labelling the 1-dimensional family of photons corresponding to energy:
  - energy
  - wavelength
  - frequency,
  - velocity (relative to a reference frequency)
  - base 10 log of any of the above
  - other transformations of the above
  - uncalibrated or indirectly calibrated instrumental channel

Basic properties

- The observable F can be photon number or energy per unit spectral parameter, optionally per area, per time, per solid angle. Examples:
  - F(nu) [erg cm^{-2} s^{-1} Hz^{-1}] vs nu, the flux density
  - nuF(nu) [erg cm^{-2} s^{-1}] vs log nu, the flux density per logarithmic interval (used in SEDs)
  - L(lambda) [erg s^{-1} A^{-1}] vs lambda, the luminosity (at the source) per unit wavelength
  - S(lambda) [erg cm^{-2} s^{-1} A^{-1} $arcsec^{-2}$] vs lambda, the surface brightness (energy per unit solid angle on the sky).
  - N(E) [photon cm^{-2} s^{-1} keV^{-1}] vs E, the number flux dennsity of photons (as opposed to the energy flux density).
- Note $\nu F(\nu) = \lambda F(\lambda) = E^2 N(E) = \nu S(\nu) d\Omega$
- Observed spectrum is corrected for sensitivity but not usually deconvolved for line spread function.

$$F(\lambda_o) = \frac{hc}{\lambda_o} \frac{N(\lambda_o)}{A(\lambda_o)\Delta t} = \int_0^\infty d\lambda F(\lambda) \frac{\lambda}{\lambda_o} A(\lambda) R(\lambda, \lambda_o)$$

Complications

- Off-diagonal line spread function (esp. X-ray domain)
- Orders in dispersed spectra (order number; overlapping order contamination)
- Different aperture sizes for different points in an SED
- Spectral artefacts
- Spatial artefacts on detector masquerading as spectral artefacts (e.g. IUE reseau marks)
- Other data quality issues
- Coordinate frames (heliocentric, rest, observed...)
- Air versus vacuum wavelengths

Proposed model

An SED is made up of a set of the following SPECTRUM objects. SPECTRUM is derived from DATA, an N-dimensional object, restricted to NAXES=1 and with a few extra attributes. The attribute names are UCD-like pending a decision on style for naming DM attributes.

- SPECTRUM - Specialization of N-dimensional data object
- SPECTRUM_NAXES: Number of coordinate axes = 1 by definition
- SPECTRUM_AXIS[1]: Coordinate Axis object
- SPECTRUM_VALUE: Observable Quantity object
- SPECTRUM_CURATION: Curation metadata object
- SPECTRUM_OBS: Observation model
- SPECTRUM_QUALITY: Overall quality object for spectrum. Should include flag for whether spectrum is fluxed or relative, or whether it is saturated, or whether it has multiple order contamination.
- SPECTRUM_EXTRACT: Extraction parameters model

The Quantity object for VALUE should have at least the following fields (which can be defaulted):

- SPECTRUM_VALUE_NAME: Observable name/UCD (FLUX_DENSITY_WAVE, NUMBER_FLUX_DENSITY_E, etc
- SPECTRUM_VALUE_UNIT: Unit, e.g. Jy, mag
- SPECTRUM_VALUE_FRAME: e.g. PHOT_SYS = JOHNSON[B] or FLUX_SYS = RELATIVE
- SPECTRUM_VALUE_VALUES: enumerated list of flux values
- SPECTRUM_VALUE_ERR_REL: enumerated list of uncertainties (or single value)
- SPECTRUM_VALUE_ERR_ABS: enumerated list of uncertainties (or single value) (INCLUDE UPPER LIMITS)
- SPECTRUM_VALUE_QUALITY: Pixel by pixel quality info
- SPECTRUM_VALUE_RES_FUNC: Line spread function array
- SPECTRUM_VALUE_QE: Sensitivity function array

Proposed model - continued

The Axis object is like the Quantity object but can specify its values via a binning scheme or a set of values:

- SPECTRUM_AXIS[1]_NAME: Spectral parameter (=WAVELENGTH, FREQ, LOG_FREQ, etc.)
- SPECTRUM_AXIS[1]_UNIT: Unit, e.g. A, cm, MHz, TeV
- SPECTRUM_AXIS[1]_FRAME: e.g. heliocentric
- SPECTRUM_AXIS[1]_SIZE: Number of samples
- SPECTRUM_AXIS[1]_VALUES: enumerated lits of bin values
- SPECTRUM_AXIS[1]_VALUES_START: left edge of bin
- SPECTRUM_AXIS[1]_VALUES_STOP: right edge of bin

In addition it may have:

- SPECTRUM_AXIS[1]_SAMPLED: data is sampled rather than binned
- SPECTRUM_AXIS[1]_START: start value of axis
- SPECTRUM_AXIS[1]_STOP: stop value of axis
- SPECTRUM_AXIS[1]_STEP: sample size in UNIT/bin; if specified, can omit VALUES

The Observation model will be specified elsewhere, of particular interest here is the Coverage sub-object SPECTRUM_OBS_COV:

- SPECTRUM_OBS_COV_SPATIAL_REGION
- SPECTRUM_OBS_COV_SPATIAL_APERTURE (approx angular size of extraction region)
- SPECTRUM_OBS_COV_TIME_START
- SPECTRUM_OBS_COV_TIME_STOP
- SPECTRUM_OBS_COV_TIME_TOTAL_EFF: Effective exposure time
- SPECTRUM_OBS_COV_WAVE_START
- SPECTRUM_OBS_COV_WAVE_STOP

The spectral-specific information are in the Extraction object: this is a placeholder for now, perhaps including:

- SPECTRUM_EXTRACT_ORDER_NO
- SPECTRUM_EXTRACT_LSLIT_ANGLE
- SPECTRUM_EXTRACT_LSLIT_ORIGIN
- SPECTRUM_EXTRACT_LSLIT_SIZE

Data Models WG - Introduction

- Mission: **Identify and standardize** the objects (software sense, i.e. concepts) in astronomy data
- Metadata is not just a list of keywords: the information is heavily **structured**.
- Most important for **data analysis** once data is retrieved
- Goals for Cambridge:
  - **Standardize process** for definining and adopting data models
  - Path from UML-style diagrams to XML implementation
  - Specific topics: 'Quantity' atomic object
  - Specific topics: 'Observation' metadata
  - Specific topics: Interoperable definition of 'Spectrum'
- WG meeting Tuesday
  - am: Introductory presentations, general discussion
  - pm: Work on collaborative document