Status of data modelling efforts

- SPACE TIME COORDS - Rots et. al.
- SPECTRUM - Will be discussed in DAL WG
- QUANTITY - Discussion on mailing list
- OBSERVATION - Summarize discussion earlier at ADASS

→

Data Model Progress outside DM WG

- Garching 2002 Jun: What's a data model? Why should I care?
- Cambridge 2003 May: OK, I guess we will need data models...
- Strasbourg 2003 Oct: Where the *!#?'s the data model?

Data Model Progress in DM WG

- Prior to this meeting: Different groups emphasise their own data models
- At this meeting: Transition to true collaboration on VO data models.
- Core team met for 7 hours Monday and Tuesday
- 2 hr WG meeting Thursday, 50 people
- Spectral DM discussion in DAL WG, Thursday
- Ready to start work on several IVOA drafts
- Face-to-face meetings planned over next few months
- Improve flow of specific requirements from other WGs

Several documents to start working on:

- DM use cases document (IVOA Note)
- OBSERVATION data model white paper (IVOA Working Draft)
- SPECTRUM data model white paper (IVOA Working Draft)

Data Model WG Summary

- Data models provide abstract description of data
- How do I unambiguously and interoperably describe my data to the VO?
- Provide basis for DAL interface design and XML metadata choices
- Cambridge meeting agreed on mini-topics to attack
- Agenda for yesterday:
- - Summary of convergence efforts for OBSERVATION models
- - Discussion on improved use cases for observation data models
- - Brief reports: QUANTITY, INTERFEROMETRY, TRANSFORMS
- - Priorization and Roadmap to XML test implementation
- - Discussion of SPECTRUM in DAL group in afternoon

Observation Models

- Scope of models vary from low level abstract 'Quantity' containers (Thomas/GSFC) to very high level views of the astronomical domain (Lemson/GAVO).
- Focus for now on observation description most needed by other WGs
- Compared models:
  - IDHA model, Louys +Bonnarel (CDS), describes image archive
  - Dataset model, Cresitello-Dittmar + CfA team, describes general dataset
  - Spectrum model, McDowell + CfA team, variation on above
  - Observation model, Giaretta (Starlink), describes general dataset
  - Canadian VO model, Dowler (CVO), describes query to archive
  - Solar observation model, Reardon (EGSO), describes solar data
- Found similarities at component level:
  - Observation, Provenance, Observatory, Coverage, Mapping, Error, Quality...
- Failure to agree on names!

Data Model Work Group Sessions Oct 14/15, 2003 at InterOpOct2003

Attendees at Oct 14 DM WG session

1. Introduction: JonathanMcDowell (JM)

Aim for the creation of a White Paper describing the 'IVOA' Data Model

- Data Model provides abstract description of data.
- Provides basis for interoperable exchange of data and queries
- Provides basis for DAL interface design and XML metadata choices
- Process:
  - White paper
  - UML model
  - XML representation - reference implementation
- DAL implements interface
- Summary of convergence efforts for Observation in various data models, which come from different points of view e.g. archive vs analysis tools. A number of data models have been considered so far, and these show areas of similarities.
  - CfA data model from JonathanMcDowell
  - Starlink one from DavidGiaretta
  - one from readon of egso
  - one from CDS- MireilleLouys?
  - others from Brian Thomas, Canadian VO, German VO *Common Areas identified:
    - Processing
    - Container
    - Instrument
    - Coverage
  - Issues: Scope and Level of detail and naming
    - Need to work on White Paper then iterate with UML diagram
    - People are encouraged to send links to DM mailing list, showing their own special data model, preferably as UML plus documentation, pointing out special features, and the mapping to common areas
      - Frank Valdes: Will send in the NOAO data model described in UML with supprting text description.
  - Common concepts:

Keywords	Concept/ Description
Dataset Observation Result	Abstract top level container for astronomical data. Could be a values of these
Provenance obtained with experiment	Abstract class describing how file was created
Observatory Observing platform "Mission"	Location where data was taken/first acquired
Coverage Meta-coverage Sample	Limits/boundaries on any parameters (possibly numeric with scientific units)
Data container Quantity NDArray	Holds information array having the same "type" - type = numeric/string/enumerated type, eventually classes - with units
Data acquisition	
Coordinate system pixel mapping Reference system Mapping WCS:Frame	Set of labels and parameters that uniquely define the context of some values
Mapping WCS:Mapping	Defines transformation of values between coordinate systems
Frameset Collection of standards (Type coords)	Contains one or more coordinate systems and zero or more mappings between them
History	The sequence/ensemble of "provenance" items
Accuracy Errors	A precise numerical measurement of the fidelity of one or more values. Example: systematic and/or statistical errors
Quality Flagged accuracy Feature	A classification of the reliability/fidelity of one or more values. The classification should contain two or more choices.

We are starting with highest level concepts; once they are agreed, we will define lower level concepts. The level of detail will be guided by the requirements for querying. The model must be extensible to allow lower level commonalities to be captured. In addition this should encourage sharing of concepts.

2. DM Use Cases: Alberto Michol

Alberto's slides in ppt format

There is a requirement for Science Cases to help in the scoping of the Data Models. Initially could look at the Science Cases from e.g. AstroGrid (AstroGrid top ten science cases) and NVO etc.

Issue of science use cases required for this.

NicholasWalton: Meeting needs to identify responsible person/people to lead the requirements capture and analysis process.

AGREED: Alberto Michol, Brian Thomas, David Giaretta, agreed to supervise this process of collation and first cut review of science requirements for Data Models. All are encouraged to send their science use cases for input into this activity.

Discussion points

Q. ??: Inclusion of plate data from smaller institutes A. JM: will be considered as a lower priority case.

Q. Javier Solano: scope to include say Planetary Data. A. JM: not first priority but will be included. Thus RA,Dec is not seen as the key identifier of objects as this isn't the case for say planetary data!

Q. Lyndsay Fletcher: assessment of quality A. Doug Tody: data characterisation is a key component of this activity, thus user assesment of 'quality' when assessing data for their use will be possible.

Reports on mini-models

Quantity - Pat D will post summary

- there is an extensive discussion on the mailing list on how to model small amount of data - quantity

Transformations: (David Berry)

- transforming values
- transforming between coordinate systems

Language was proposed on mailing list involving abstract classes and subclasses, including compound mappings so that arbitrarily complex mappings. DT proposed that such transformations are the remit of analysis software; it was decided to let work be driven by requirements. Additional ideas can be found in Radio data - in AIPS++ MeasurementSet?. - ALMA data format and associated model

3. Road Map

Requirements needed by the DAL group. Doug Tody noted upcoming input required from the data model needed by the DAL group in their v1.0 SSA and v1.1 SIA standards.

Short term goals: Priority:

- Spectral Data Model for SSA
- Observation data model
- Incremental "deliveries" are needed in order to have useable pieces
- Consider Packaging as a possible area of work
  - Mosaic images
    - There are several different ways of packaging these currently

Other issues:

- Data Collections - e.g. from an IFU - maybe just
- Updating of data infomation - maybe a registry problem (e.g. handle update of photmetric calibration data)

Brian Thomas: would like to see a more developed framework for the development of the Data Model

On the whiteboard:

- slow> top level: obs, spectra etc
- intermediate> mid level: components
- fast> low level: qualntities, units

(There was disagreement on which activities were slow and which were fast)

Move with idea of local namespaces.

JonathanMcDowell: Names against Milestone Tasks

- White Paper at Observation and Component Level - end Dec 2003
  - Take a Use Case to implement and pass to the DAL group for consideration
- UML Level Feb 2003.

Francois Bonnarel - common rules - ensure compatability with DAL.

-- NicholasWalton & DavidGiaretta - 16 Oct 2003

Revision r1.2 - 16 Oct 2003 - 16:04 GMT - DavidGiaretta
Parents: WebHome > IvoaInteroperability > IvoaDataModel

Quantity

↑ ← →

- Basic idea of QUANTITY included values, errors, units
- Many different versions presented
- Most included a simple subobject similar to Ray's initial suggestion
- Need for a more complex MEASUREMENT object which includes coord frame, resolution, exposure subobjects
- Issue: Should simple object be instance of complicated with other things set null, or should complicated object inherit from simple object?
- In July: presented attempt to synthesize discussion
- Some reluctance to compromise

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.
- We distinguish between a spectrum in the theoretical sense, the energy output versus e.g. frequency $F(\nu)$, and a {\bf 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 metadata 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

- Other observables as a function of wavelength: percentage polarization, extinction coefficient. These can use the present model.
- 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.
- Spectral coordinates for particles other than photons: massless (gravitational waves) or massive (electron energy dist. in radio jet, cosmic ray spectrum).
- Spectral coordinates not a particle property: power spectra of source variability or CMB anisotropies, Fourier transforms in general. Needs a slightly different model.

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}~$\AA^{-1}$
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} $\AA^{-1}$
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} $\AA^{-1}$ 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

↑ ← →

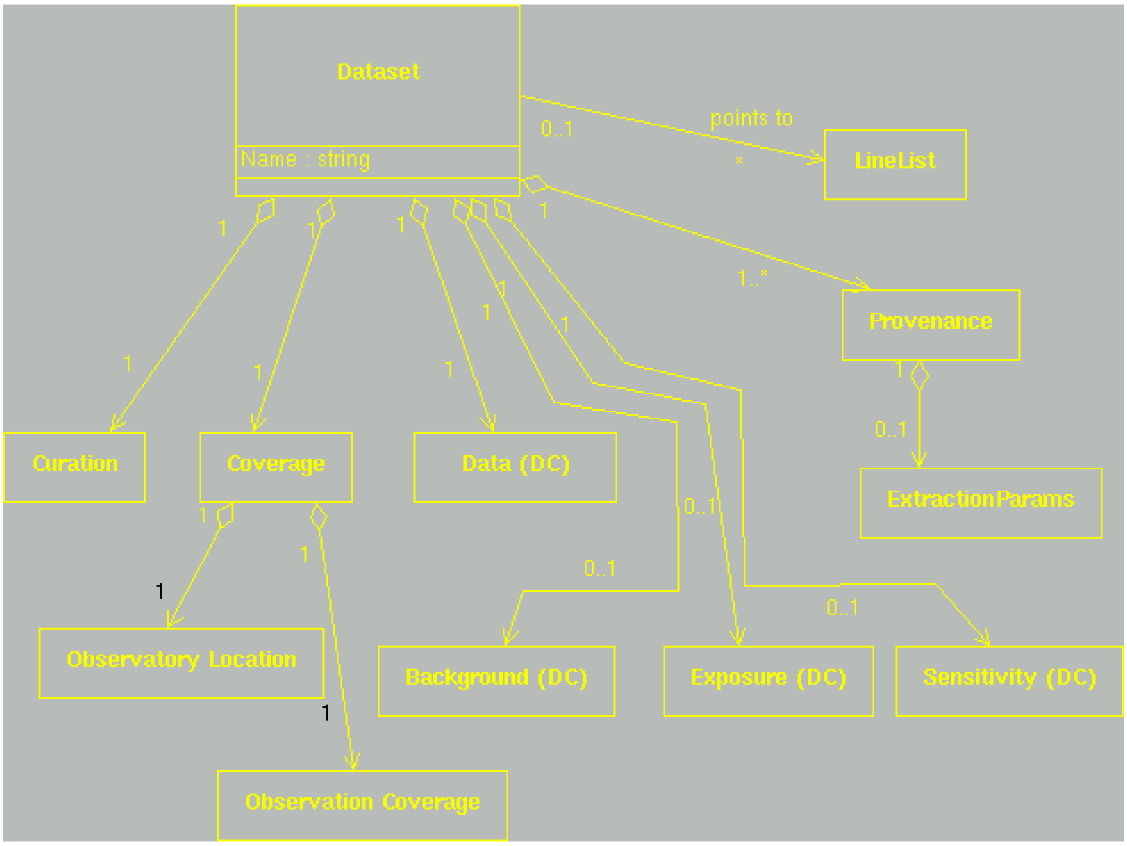

Fig. 1: The complete Dataset object, which contains curation and coverage objects as well as several Data Container objects. The dataset will have at least one Data Container for the main data, and may have additional ones for a background spectrum, an exposure array, and a sensitivity array.

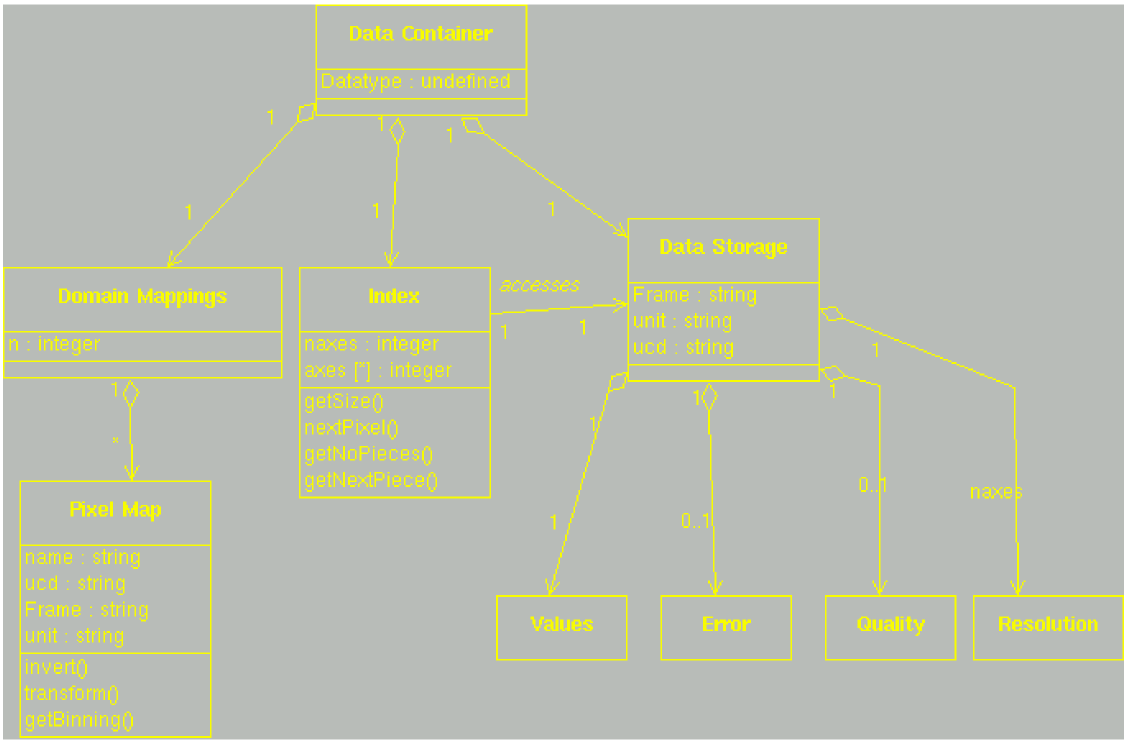

Fig. 2: The Data Container has a Data Storage object containing Value, Error, Quality and Resolution sub-objects.

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.

Fig. 3: The Observation Coverage object is a summary of the Observation Location object described in the Space Time Metadata of Rots et al. (hea-www.harvard.edu/$\sim$arots/nvometa) and encapsulates the spatial and temporal region from which the spectrum was extracted. Observation Coverage ({\it what direction was I looking in?}) together with an Observatory Location object ({\it where was I standing at the time?}) provides software the context of an observation in space, time, and frequency.

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. Alternative 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.

Special cases from use survey

- Antenna temperature spectra (e.g. SWAS) - include in possible observables
- Wavenumber spectra *do* exist in archives - e.g. Arcturus atlas
- Ratio spectra, e.g. Arcturus/Telluric - include in possible observables; needs extra metadata?
- Combination spectra using same X axis: e.g. Arcturus, Telluric, Ratio. Design decision: require them to be presented as three different spectra? Or implement array of spectra (note observables not all the same)
- May need to include object redshift, even if data presented in observed frame
- Corrections to observable: absorption (telluric, galactic, etc), continuum subtraction
- Other flags to observable: fit, model, etc.
- Line IDs and parameters - a separate problem out of this scope?
- Arc lamp spectra from NOAO - what metadata?
- SDSS spectra: nx4 image, with different observables in each plane (data, continuum-subtracted, error, mask).
- Other metadata: spectral type, etc. - at higher level index?

Summary

⟶

- Data models provide abstract description of data
- Provide framework for interoperable exchange of data and queries
- Provide basis for DAL interface design and XML metadata choices
- Cambridge meeting agreed on mini-topics to attack
- Initial discussions on QUANTITY and SPECTRUM objects
- Now returning to focus on larger scope OBSERVATION
- Agenda for today:
- - Summary of convergence efforts for OBSERVATION models
- - Discussion on use cases for observation data models
- - Brief reports: QUANTITY, INTERFEROMETRY, TRANSFORMS
- - Priorization and Roadmap to XML test implementation
- - Discussion of SPECTRUM in DAL group this afternoon