

## ACCESS TO SPECTROSCOPIC DATA IN THE VO

**Douglas Tody**

National Radio Astronomy Observatory

US National Virtual Observatory

### ABSTRACT

The Virtual Observatory (VO) provides client applications with a uniform interface to access distributed, multi-wavelength astronomical data. In this paper we review the VO data access interfaces, and describe the capabilities provided for accessing spectroscopic data, including simple one-dimensional spectra, spectral aggregates and SEDs, spectral data cubes, and observed and theoretical spectral line lists. These capabilities include comprehensive metadata for uniformly describing and characterising datasets, an essential capability for enabling automated data discovery and data selection, and capabilities to dynamically subset and filter data at access time, or actively mediate external data to a common data model, an essential capability for automated analysis of data from many sources.

Key words: Technique: spectroscopic; Virtual Observatory.

### 1. INTRODUCTION

Modern astronomical data analysis often requires use of data from multiple instruments or data collections, combining data from multiple branches of astronomy. Data from custom observing programs may be combined with data from public archives, or with data from theoretical simulations. The data to be used for a given analysis may be selected from collections containing millions of individual datasets. The individual datasets to be accessed may be very large, possibly gigabytes or terabytes in size, and typically are stored remotely. While data from a single instrument or survey may be homogeneous, data from multiple sources is often heterogeneous and idiosyncratic.

While it has long been possible to find and download data from multiple sources and perform multi-wavelength data analysis, the process is generally interactive and can be very time consuming, limiting the size of the problems

which can be addressed, and the amount of science which can be accomplished. At the same time the volume of data produced by newer instruments and large scale surveys is growing exponentially, and it will be difficult or impossible to continue to utilise future data effectively with the techniques used in the past. A major goal of VO is to automate the process of discovering and accessing multi-wavelength astronomical data, so that analysis can scale up to deal with the volume and diversity of the data now becoming available.

The VO data services address these problems, providing client applications with a uniform interface to data regardless of how or where it is stored, and regardless of how the data originated. Standard metadata is provided to provide a uniform view of data regardless of its origin, enabling automated data discovery and selection. Support is provided for dynamic sub-setting, filtering, and transformation of external data, allowing very large datasets to be accessed efficiently by a remote application. Access may involve on-the-fly transformation to a standard data model or data format, so that client applications do not have to deal with the details of how data is stored or represented externally within each project publishing data to the VO.

The VO services specific to spectral data are part of a broader suite of services providing access to astronomical data in general. To understand what is provided we will look first at the general problem of science data access in the VO, including the role data services play and how they function, followed by a closer look at the capabilities provided specifically for spectral data.

### 2. DATA SERVICES IN THE VO

#### 2.1. Architecture

The most important thing to understand about data access in the VO is that it is all about data services or "middleware". This is a new layer of software between the client application, and the data being accessed. The data access

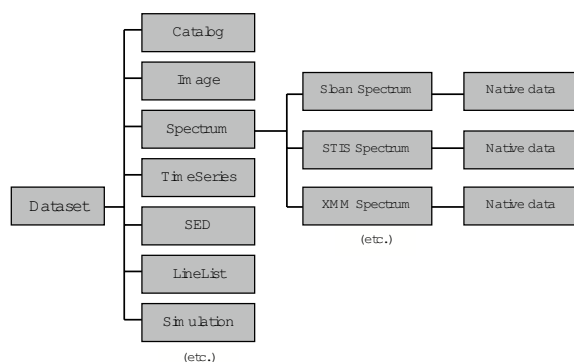


Figure 1. Astronomical data in the VO is treated as a class hierarchy, with the generic dataset at the root, and successively more specific types of data below, until we reach individual instrumental or survey data collections.

layer provides *location transparency*, so that the client application does not need to worry about where the data is stored, or whether it is local or remote, and a *uniform interface* to the data, so that the client application does not need to support a custom interface to each individual data archive or repository, or for each mode of data access. The data access layer also provides a *uniform view* of heterogeneous data, by defining standard dataset metadata, and in some cases standard data models for representing actual data.

In most cases the VO does not provide the client applications used by an astronomer, nor the astronomical data which these applications access. Most of these instead come directly from the astronomical community. Hence we assume that existing applications will need to become “VO aware”, by modifying these applications to use some VO client library which allows them to talk to the VO and access data via the VO framework and services. Likewise, existing data from astronomical archives is *published* to the VO by implementing a data service which understands the details of the data, and can provide a VO-compliant view of the data. The data itself as stored externally does not have to be modified.

## 2.2. Classes of Data

Astronomical data in the VO is treated as a class hierarchy, with the “generic dataset” at the root. Beneath the generic dataset we have the major classes of astronomical data, *catalogue*, *image*, *spectrum*, *time series*, and so forth. Beneath each of these we have data from specific data collections. This hierarchy is illustrated in figure 1.

The significance of a class hierarchy such as this is that any facilities we provide for a generic dataset can be applied to any catalogue, image, spectrum, or other class of data; likewise, any facilities we provide for a generic spectrum can be applied to an actual spectrum from a specific data collection. Any type of astronomical data may be viewed as a generic “dataset”; any type of “spectrum”

may be viewed either in a generic sense, or with knowledge of any collection-specific metadata or other features defined for the specific data collection.

Without an approach such as this it would be very difficult to provide a uniform view of data, and uniform access interfaces. By sub-classing data we are also able to provide the collection-specific details required to fully understand data from a specific instrumental, survey, or theory data collection.

## 2.3. Solar and Planetary Data

While the VO generally favours Galactic and extragalactic data, the approach taken is general enough to be useful for much solar and planetary data as well. The concepts of “table”, “image”, “spectrum”, etc., are general enough to apply to solar and planetary data as well as to non-solar system data. Much of the metadata defined is also general enough to apply to solar and planetary data. For example, metadata for dataset identification and curation, and for physical dataset characterisation, are quite general. In general, any coordinate system may be used, including those defined for solar and planetary data.

## 3. DATA SERVICE FUNCTIONALITY

### 3.1. Data Discovery

Data access in the VO typically begins with a call to the global VO *registry* to find data services of interest. Usually, these are services which serve data relevant to whatever analysis is being performed, e.g., based on waveband, or position on the sky. Often many relevant services will be found.

The next step is for the client application to issue a query to each data service, specifying in more detail the data of interest. Since we have a uniform interface, the *same query* can normally be submitted to all such services. If desired, these queries can all execute concurrently.

Each service will then return a query response, which is a table, in VOTable (XML) format, listing all the datasets available from that service which match the query. Since many such services may be queried, this provides a *global data discovery* capability. New data services may come on-line at any time, and the data services are queried directly to discover data, hence data discovery will find new or modified data as soon as it becomes available.

### 3.2. Metadata Access

The query response returned by a data service lists each available dataset matching the query. For each such

dataset, standard *metadata* is returned describing the dataset in some detail. Often all the client requires is this metadata, and the data itself may not be accessed. If data is to be accessed, the detailed query response metadata will often be used to select only some fraction of the available datasets for download.

### 3.3. Data Access

Included in the query response metadata describing a dataset is an *access reference* URL which can be used to retrieve the data. In the simplest case, data retrieval is very straightforward, using a protocol such as HTTP or FTP to synchronously download the data. If desired, multiple datasets can be downloaded concurrently.

Often the dataset described in a query response does not physically exist at query time; it is a *virtual data*, which is computed on the fly if and when it is actually accessed. This capability is essential for solving certain problems in the VO, for example access to very large datasets, and access to heterogeneous data.

In the case of very large datasets, the "dataset" to be accessed may be a small subset of the physical dataset as stored in the remote archive. For example, a spectrum may be returned containing a cutout around a single line of a large high resolution spectrogram, or a spectral data cube may be returned containing a sub-cube of a much larger data cube.

In the case of heterogeneous data, the data returned to the client application may be converted on the fly into a standard form, regardless of how data is physically stored or represented in the remote archive. Virtual data generation also makes it possible to view the same data in multiple ways, for example by dynamically extracting a 1-D spectrum from a spectral data cube (which is an image).

## 4. METADATA AND DATA MODELS

### 4.1. Dataset Metadata

Dataset metadata describes a dataset, e.g., a single spectrum or image. An example from the pre-VO days would be a FITS image header. In the case of VO we require a uniform view of data, hence a lot of effort has gone into defining standard dataset metadata.

Dataset metadata comes in two main varieties, *generic dataset metadata*, which is valid for any type of data, and *type-specific metadata*, which can vary as necessary to describe each major class of astronomical data (image, spectrum, time series, etc.). In addition, the metadata mechanism is extensible, allowing custom metadata specific to a given data collection to be added to convey detailed information specific to the data, e.g., the instrument configuration.

Table 1. Types of generic dataset metadata.

Type	Description
DataID	Dataset identification (title, creator, etc.)
Curation	Dataset curation (publisher etc.)
Target	Observed astronomical target, if any
Derived	Derived quantities (SNR, redshift, etc.)
CoordSys	Coordinate systems and reference frames
Characterisation	Physical characteristics of axes

The major classes of generic dataset metadata are shown in table 1. The Characterisation model describes the physical characteristics of the data, including the coverage, resolution, sampling, and errors associated with each measurement axis (spatial, spectral, temporal). In addition, the query response may also return information describing logical associations of related datasets, and how to access the data.

### 4.2. Spectral Data Model

While standard metadata can be used to describe a dataset, a *data model* is required to provide a standard way to represent the dataset itself. More precisely, a data model specifies the *semantic content* of a dataset. The same content can be represented in multiple physical formats, e.g., VOTable, FITS, or native XML.

Accessing 1-D spectra in the VO is especially problematic as, unlike the case with tables or images, there is no standard way within astronomy to represent spectra. As a result, nearly every spectral data collection has its own unique project-specific format for spectra. It is infeasible for a client application to understand the details of how spectra are represented in every project-specific spectral data collection within astronomy, so VO has defined a standard *Spectrum Data Model*, including standard representations for VOTable and FITS.

Although the current Spectrum data model deals explicitly with simple 1-D spectra, the model is more general than that, and will eventually be extended to apply to all spectral or spectrophotometric data, including 1-D spectra, SEDs, and the spectral axis of spectral data cubes, as well as related spectrophotometric data such as time series and the photometric flux of pixels in a 2-D image.

## 5. VO SERVICES FOR SPECTROSCOPIC DATA

The VO data access protocols for spectroscopic data are still under active development. The protocols for access-

ing 1-D spectra and spectral line lists are currently the most advanced, including implementations.

### 5.1. One Dimensional Spectra

The case of one-dimensional spectra is handled specially to provide a simple, optimised solution for this very common case. Much data from large spectral surveys is of this form, plus more complex spectra can often be viewed as 1-D spectra, or as an aggregation of such spectra.

In VO, the *Simple Spectral Access* interface (SSA) is provided to access 1-D spectra. A query interface is provided which allows discovery queries based on position on the sky, spectral bandpass, time coverage, and various other attributes such as spatial and spectral resolution, signal-to-noise ratio, redshift, variability, and so forth.

Both calibrated and uncalibrated data can be accessed. Full support for error estimation is provided. In the most general case, most measurement attributes, such as the bin size, resolution, errors, quality, etc., can be specified separately for each sample point. Fluxes can be specified in a variety of units including absolute flux and photometric magnitudes.

Both observed spectra and synthetic (model-based) spectra can be accessed. Theoretical spectral can be generated on the fly if desired, passing custom parameters to drive the model. Dynamic extraction of spectra from more fundamental data (such as a data cube) is supported. Spectra can be returned in VOTable, FITS binary table, native XML, or CSV format, compliant to the Spectrum data model, or native project spectra can be passed through. Graphical or HTML renditions can also be returned.

### 5.2. Spectral Aggregates

One-dimensional spectra are often related in some fashion, e.g., an Echelle spectrogram may be represented as an aggregation of individual 1-D spectra, or data from a discrete IFU instrument may be viewed as an aggregation of 1-D spectra, all taken at the same time and with the same observing parameters. These are cases of *complex datasets* which are composed of simpler primary datasets, such as individual 1-D spectra.

We are still discussing how best to model such data in the VO, to allow complex aggregates to be described without overly complicating the software. Probably there is no one best solution. Generalising the 1-D spectrum case to support multiple segments is one solution. Another approach is to relate spectra in the discovery query to allow complex aggregates to be expressed while still basing access on individual simple 1-D spectra. Probably the best solution is a combination of the two approaches.

### 5.3. Spectral Energy Distributions (SEDs)

In recent years SEDs have become a major tool for astrophysical research. They also represent a challenging use-case for VO, due to the need to combine data from multiple multi-wavelength observations in a physically meaningful fashion, taking into account the quality of flux or other calibrations, possible source confusion due to variable resolution or object type, and many other factors.

How best to handle SEDs in the VO is still an active topic for discussion. One approach is to define a SED as a new type of astronomical data, similar to a spectrum but very different in the details of what is presented. This would allow data centres, or automated SED generation tools, to publish ready-made SEDs to the VO. Another approach is to provide only tools for generating SEDs, leaving it up to individual researchers to generate their own SEDs as part of their research program. Even in the latter case, if these custom SEDs are referenced in a published paper, one might still want to be able to access the digital data.

Probably both approaches are needed. Ready-made SEDs can be useful for a quick look at the spectral energy distribution of a source, even if more careful analysis is needed. Tools for manual or automated SED generation are needed in all cases.

### 5.4. Time Series Data

Light curves are not normally considered spectral data, but if we look at light curves and 1-D spectra from a data modelling perspective, there is very little difference between them; in fact the same model and access interface can be used to represent both. Both are sequences of spectrophotometric samples, which may be irregularly spaced or segmented, with associated general dataset and observation metadata. The distinction can be further blurred by cases such as a time series of spectra, for example from synoptic surveys, radio GRB surveys, or solar data.

For a spectrum, the spectral coordinate is the primary variable; for a time series, time is the primary variable. However in both cases, time or spectral coordinate can be a secondary variable, for example in the case of a time series where, at each sample point, photometry is obtained simultaneously through several standard filters or bandpasses.

### 5.5. Spectral Data Cubes

In VO currently, access to 2-D images is provided by the *Simple Image Access* interface (SIA). Version 2.0 of this interface, currently in the planning stages, will support multi-dimensional data, primarily spectral data cubes and time cubes. Although in general such data may be multi-dimensional, for simplicity we refer to it as “cube” data,

and indeed the two and three dimensional cases are the most important in terms of representing real astronomical data.

While modelling and representation of cube data is non-trivial, it is in some sense a straightforward generalisation of what has already been done for one and two dimensional data (spectra, time series, and 2-D images). Nearly all of the data model elements and metadata can be reused for cube data. Related work such as FITS WCS is also very relevant.

The real challenge with cube data is data access, which can be considerably more complex for cube data. Typical access modes for a cube include the following:

- The entire cube.
- Any 3-D sub-cube (no resampling).
- Any 2-D plane (no resampling).
- 3-D projection onto a 2-D plane.
- Standard 2-D renditions, e.g., continuum.
- General 2-D slice at any 3-D position or orientation.
- 2-D or 3-D reprojection of a subregion (warping).
- 1-D spectral extraction with a user-specific aperture.

Operations such as projection may optionally include filtering of data along the axis to be collapsed; in the case of a spectral data cube for example, this could be used to remove night sky lines.

Flexible access to data cubes such as we describe above is extremely important for remote access to cube data, as data cubes can be very large. For example, a 2K by 2K cube with 8K spectral channels, at 4 bytes per sample, is 128 GB in size. If we add full polarisation, a single cube dataset can reach one half Terabyte in size! This is beyond anything yet far attempted, but is not at all unreasonable for future instrumentation now being built.

## 5.6. Spectral Line Lists

Analysis of spectral data often requires access to spectral line data, including both laboratory data and data from theoretical models. To support this, VO provides both a physical data model for characterising spectral lines, and a *Simple Line Access Protocol* interface (SLAP), for access to spectral line data. In the case of access to data from theoretical models, it is possible to pass custom model parameters to the service, so that line data can be generated on the fly, either from the model or from a large database of pre-generated model data.

## 6. SUMMARY

The virtual observatory defines a middle layer between the client applications and data already provided by the astronomical community, which provides uniform access to local and remote data from all branches of astronomy. Comprehensive, standard metadata and data models, as well as access methods, are provided to uniformly describe and represent astronomical data, enabling automated selection and processing of astronomical data. In the case of spectroscopic data, support is provided both for data which is explicitly spectroscopic, such as 1-D spectra, but the more subtle case of a spectroscopic or spectrophotometric axis associated with other forms of astronomical data such as images or time series, is also fully supported.

## ACKNOWLEDGMENTS

Many people have contributed to the development of the work described in this paper. In particular Jonathan McDowell, Markus Dolensky, our other co-authors of the related IVOA specifications, and the members of the IVOA Data Access Layer working group, helped develop the concepts presented here in many working group meetings and discussions. This work was supported via a grant from the National Science Foundation's Information Technology Research program to develop the US National Virtual Observatory, as well as via various IVOA partners, including the Framework Programmes of the European Community.

## REFERENCES

- Greisen, E., Calabretta, M., Valdes, F., Allen, S.L, 2005, *A&A*, 446, 747.
- Louys, M., Richards, A., Bonnarel, F., Micol, A., Chilingarian, I., McDowell, J., *Characterization Data Model*, 2007, IVOA Proposed Recommendation.
- McDowell, J., Tody, D., et al., *Spectral Data Model*, 2007, IVOA Proposed Recommendation.
- Ochsenbein, F., et al., *VOTable Format Definition*, 2004, IVOA Recommendation.
- Rots, A. *Space Time Coordinate Metadata for the Virtual Observatory*, 2007, IVOA Proposed Recommendation.
- Salgado, J., Osuna, P., et al., *Simple Line Access Protocol*, 2007, IVOA Working Draft.
- Tody, D., Dolensky, M., et al., *Simple Spectral Access Protocol*, 2007, IVOA Proposed Recommendation.
- Tody, D., Plante, R., *Simple Image Access Specification*, 2004, IVOA Working Draft.

This page is left intentionally blank.