*I*nternational

*V*irtual

*O*bservatory

*A*lliance

Time Series Data in the VO

Version 0.1

IVOA Note 2010-12-02

This version:

ThisVersion-YYYYMMDD

Latest version:

http://www.ivoa.net/Documents/latest/latest-version-name

Previous version(s):

Authors:

TBD (D.Tody compiled this first draft to get things started)

Contributors:

TBD

…

Abstract

A draft standard data model and access protocol for time series data is proposed, building on work already done in the IVOA to represent and access photometric data and spectra as well as experience with time series data within the broader astronomical community. This is intended only as a starting point to develop the necessary standards for time series data, leading eventually to actual IVOA working drafts for the required time series data model and access protocols.

Status of This Document

This is an IVOA Note. The first release of this document was 2010-12-02.

This is an IVOA Note expressing suggestions from and opinions of the authors. It is intended to share best practices, possible approaches, or other perspectives on interoperability with the Virtual Observatory. It should not be referenced or otherwise interpreted as a standard specification.

A list of current IVOA Recommendations and other technical documents can be found at http://www.ivoa.net/Documents/.

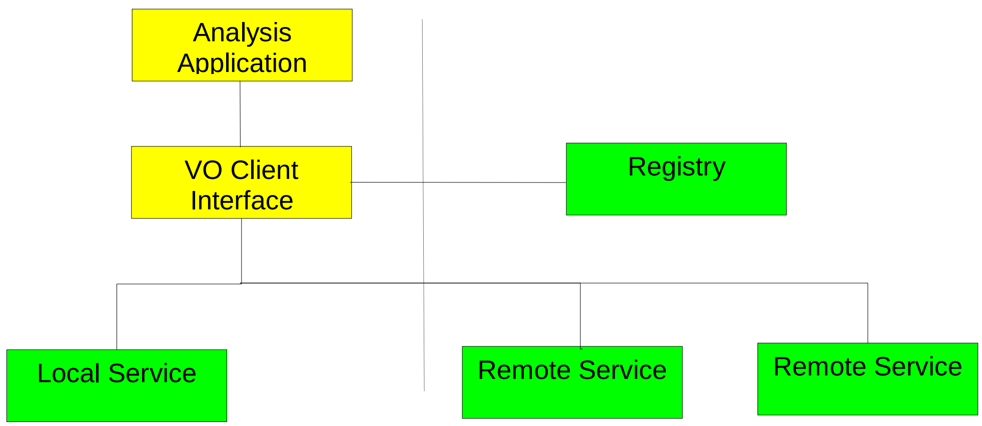
Acknowledgements

"Ack here, if any"

Contents

1 Introduction

This document outlines a strategy for how to do distributed analysis of archival time series data in the virtual observatory (VO). The class of time series data dealt with here consists of a sequence of discrete values represented as a function of time. A typical example would be a light curve measuring flux as a function of time for a single astronomical object, possibly in several colors or photometric bands, but in general any quantity could be the dependent variable. We do not deal here with other types of time



domain data such as transients or synoptic images and spectra.

The architecture of a typical VO-based time series analysis application is shown in the figure above.

In the VO a client application typically accesses data remotely, communicating with a data service on the remote server which hosts the data. A standard *data access protocol* is used to communicate with the remote service. The VO *resource registry* may be used to discover relevant data services. Data services may also be colocated on the same local area network as the client application; this is transparent to the client application but in this case the network connection to the service will have a much higher bandwidth. The range of possible client applications is completely open ended, e.g., a desktop client analysis application, a Web browser, a Web-client application located on another server and accessed remotely by the user, a workflow, and so forth. All share the same standard VO interfaces for service discovery and data access.

2 VAO applications

[To be added] Discuss:

- *Scientific capabiities to be provided*

- *Implementation in terms of VO architecture*
- *Any unusual requirements on VO infrastructure (for example should we provide access to derived data products such as a periodogram as well as the photometry data within the access protocol, as opposed to via a separate custom analysis service/application? Anything else like that which is desirable at this level?).*

3 Time Series Interface

The international VO community in recent years has developed much of the technology required to represent and access astronomical data in the VO. Data access requires both a *data model* to describe the specific type of data being accessed, and an *access protocol* defining how to access such data in the VO. The different classes of data such as images, spectra, and time series have many characteristics in common hence their data models and access protocols are similar, while still reflecting the special characteristics and analysis requirements of the particular class of data being accessed. In particular, much of the metadata required to describe and characterize an astronomical dataset is common to all these interfaces. Basic features of the data access protocol such as searching by object name or class, position, or spectral or time coverage are common to all the data access interfaces.

While most of the basic features of the data model and access protocol described in the following sections are generic and common to most astronomical data, some aspects have been extended to support time series data. In particular the photometry model has been enhanced to better describe photometric bandpasses and facilitate conversion to absolute flux units, and support has been added to describe time series data where measurements are made simultaneously (or nearly so) in multiple bandpasses. [*These new features, in particular the Photometry model, are still under active discussion within the IVOA community].*

3.1 Data Model

Time series data and 1D spectra are closely related: a time series deals with flux vs time whereas a spectrum deals with flux vs spectral coordinate. Both can be represented as special cases of a more general spectrophotomeric model. The IVOA has already defined a standard data model for spectra (the Spectrum data model [*ref*]) and associated access protocol (Simple Spectral Access Protocol or SSAP [*ref*]). Most of what is described herein is taken directly from the Spectrum data model and SSA interface. We only summarize the major elements of these models here; many details are omitted here and the Spectrum documents should be referred to for a more complete specification of the data model and interface. Where extensions are required for time series data we explicitly discuss them here.

The time series data model is intended to be able to describe a wide range of time series data. Light curves of observed objects are the most common example, however in general any observed or derived quantity may vary with time and constitute a valid time series dataset.

The time series data model is used both to *describe* time series datasets and to *compose* time series datasets. We can use the data model both to describe datasets (individual data products) available for retrieval, as well as to compose the actual datasets returned to the client. Passing native project data through unchanged to the client is also possible. The same time series data content can be returned in a variety of data formats, e.g., FITS, VOTable, XML, HTML, text, and so forth.

The following data model components are used to both describe and compose time series datasets:

Component	*Description*
Access	Used by the access protocol to provided metadata required to access or retrieve the dataset.
Association	Used in the query response of the access protocol to associate datasets which are associated in some way.
Dataset	Catch-all used to describe attributes of the dataset which are not covered by the more generic attributes of the model.
DataID	Dataset identification. General information about this specific dataset, e.g., the title, the creator, the unique dataset identifier, and the instrumental, survey, or other data collection to which the dataset belongs.
Curation	Information about the archive or data center responsible for publishing and curating the dataset.
Target	Information about the observed astronomical target if any, e.g., name, description, and classification.
Derived	Any quantities derived from analysis of the dataset, e.g., percent variability or redshift.
CoordSys	Defines the default coordinate system frames used to characterize the data or supply actual observed data values. May be overridden for individual data elements.
Characterization	Characterizes the physical attributes of the overall dataset, e.g., location, coverage, resolution, etc. of the spatial, spectral, time, and sometimes polarization axes.

Data Values	The actual data points comprising an observed or computed dataset; omitted when the dataset is merely described, e.g., in a discovery query.

Each of these components (little data models) may contain any number of attributes. Some of the more important ones are summarized in the following sections. A more comprehensive list is given in [*ref to DM spreadsheet or other documentation*].

Other data model components are possible. For example a *provenance* data model is currently being defined which will allow more to be said about the origins of a particular dataset, e.g., the configuration of the instrument or model which produced the data, or a description of the processing performed to compute the dataset. The data model is exensible, allowing a data provider to provide additional non-standard metadata to describe the attributes of a specific data collection.

These data model elements are specified in more detail in the following sections. Most of this is standard VO dataset metadata which we merely summarize here. Elements which have been added or modified to support time series data are highlighted in bold text.

It should be noted that much of this metadata is optional; nonetheless it is useful to define standard metadata for cases where it is needed and can be provided. The subset of this metadata which should be considered mandatory for time series data is TBD.

3.1.1 Access Metadata

Access metadata is used to tell a client how to access the datasets described in a query response describing available data matching the client's search criteria.

Access.Reference	URI (URL) used to access the dataset
Access.Format	MIME type of the dataset (data file)
Access.Size	Estimated/approximate dataset size

3.1.2 Association Metadata

Association metadata is used to describe logical associations relating datasets described in the query response, that is, datasets which are related in some fashion. This metadata is not used in an actual time series dataset, only in the query response of a data discovery query.

Association.Type	Type of association (defined externally)

Association.ID	Unique ID identifying the association
Association.Key	Unique key indicating what is different for each element of the association.

A simple example of an association is a *multiformat* association where the same dataset is available in several different output formats or views. Another example specific to time series could relate a time series and its associated periodogram, with a service able to return either.

3.1.3 Dataset Metadata

Dataset metadata describes the overall "dataset" (in this context *dataset* refers to a single data object, normally represented as a file of some sort, e.g. an individual time series dataset).

Dataset.DataModel	Data model name and version.
Dataset.Type	Dataset or segment type, e.g., TimeSeries
Dataset.Length	Number of data points
Dataset.Period	Period of time series if relevant/known
Dataset.TimeAxisIsFolded	Time axis values are folded by the period

The *Dataset.Period* and *Dataset.TimeAxisIsFolded* attributes have been added here and are specific to time series data. Some standard Dataset attributes are not shown here, e.g. SI scaling factors for the data vectors.

3.1.4 Dataset Identification Metadata

Dataset identification metadata is used to describe the fundamental identity of a dataset, including where it came from and how it was created.

DataID.Title	Dataset descriptive title (free form)
DataID.Creator	Creator name (string)
DataID.Collection	IVOA identifier of data collection (string)
DataID.DatasetID	IVOA dataset identifier, e.g., ADS id
DataID.CreatorDID	Creator assigned dataset identifier
DataID.Date	Data processing/creation date
DataID.Version	Version of creator-produced dataset
DataID.Instrument	Instrument name (if any)
DataID.Bandpass	Bandpass name or names, e.g. filter
DataID.DataSource	Original source of data (e.g. survey)
DataID.CreationType	Dataset creation type

All of this is standard VO metadata with the exception that *DataID.Bandpass* is modified to permit a list of bandpass (e.g. filter) names. Note there is as yet no standard for naming bandpasses or filters, so the contents of this field if provided are up to the data provider.

3.1.5 Curation Metadata

Curation metadata describes who curates the dataset and how it is provided to the VO, e.g., the project, data center, or archive which curates the data and makes it available online.

Curation.Publisher	Publisher name (string)
Curation.Reference	URL or Bibcode for documentation
Curation.PublisherDID	Publisher assigned dataset identifier
Curation.Date	Date curated dataset last modified
Curation.Version	Version of curated dataset
Curation.Rights	Restrictions: public, proprietary, etc.

All of this is standard VO metadata. The *PublisherDID* is especially useful as it uniquely identifies this particular dataset within the context of the given publisher (archive), allowing reliable retrieval once the PublisherDID is known.

3.1.6 Target Metadata

Target metadata describes the astronomical target observed, if any.

Target.Name	Target (astronomical object) name
Target.Class	Target or object class
Target.Redshift	Target redshift
Target.VarAmpl	Target variability amplitude (0-1)
Derived.SNR	Signal to noise ratio

The target class is the type of astronomical object, if known, e.g. AGN, cepheid, blazar, GRB, supernovae, etc., such as might be known apriori or determined by a classification algorithm [*for a classification algorithm we probably need additional metadata to describe the classification in probabilistic terms.*]. There is as yet no standard formally defined classification scheme; informal keyword-based classification such as is used in publications can nonetheless be quite useful. VO semantic analysis techniques based upon ontologies could provide a capability to search for data based upon classification.

The variability amplitude is a standard value but especially important for time series data. An object with a variabillty amplitude of zero is not considered variable and essentially has no light curve (or a flat light curve within the measurement error).

Derived.SNR is actually a derived quantity as defined in the next section, but is included here for simplicity and to suggest what to include in a query response.

3.1.7 Derived Quantities Metadata

A distinction is made between *target* metadata, defining standard attributes for a given astronomical target, and *derived* metadata, which defines values actually derived from the particular dataset being described. For example the variability amplitude for an object may be defined in some standard catalog, but may differ from the derived value for the actual dataset (observation) being described. While VO (the Spectrum data model) makes this distinction, it is not clear if it is practical to provide both values in actual data collections. A pragmatic solution is to provide only the values defined for TARGET above, while allowing additional DERIVED values to be provided if known. If no "well known" TARGET value is available, the derived value should be used instead for the target model.

3.1.8 Coordinate System Metadata

Coordinate system metadata defines the default coordinate systems and reference frames used for both data characterization and data values. These may be overridden locally in either context.

CoordSys.ID	Unique ID string for coordinate system
CoordSys.SpaceFrame.Name	Spatial coordinate system name/type
CoordSys.SpaceFrame.Ucd	UCD of spatial frame
CoordSys.SpaceFrame.RefPos	Origin of spatial frame
CoordSys.SpaceFrame.Equinox	Equinox of spatial frame (if used)
CoordSys.TimeFrame.Name	Timescale
CoordSys.TimeFrame.Ucd	UCD of time frame
CoordSys.TimeFrame.Zero	Zero point of timescale in MJD
CoordSys.TimeFrame.RefPos	Location of times for photon arrival
CoordSys.SpectralFrame.Name	Spectral frame name
CoordSys.SpectralFrame.Ucd	UCD of spectral frame
CoordSys.SpectralFrame.RefPos	Spectral frame origin

The *TimeFrame* values above are standard (defined for any type of data) but are especially important for time series data [*do we need anything else here?*].

It is not clear how important the *SpectralFrame* values are for time series data since we often explictly define the spectral (photometric) band used, however we include them here for completeness. For some data, e.g., radio or high energy, spectral bands (in these sense of filters in the O/IR) are not used and these values may be required.

3.1.8.1 Spectral Bands

Photometric systems are especially important for much time series data, which is often produced by observation through a filter or other stamdard-bandpass instrumental mechanism. Furthermore time series are often produced simultaneously (or nearly so) in several spectral bands. A photometric system can be considered as a type of coordinate system mapping "world coordinate" flux values of some sort (e.g., magnitudes) to an absolute flux value. Hence we propose ([*ref Photometry data model*]) to describe the photometric bands used in a time series as an element of CoordSys.

At a minimum the spectral bands observed should be described sufficiently to identify the individual flux measurements in a time series. In the more general case sufficient metadata to compute calibrated flux values may be needed, e.g. to use time series photometry as input to compute a spectral energy distribution.

The proposed photometry model currently defines the following attributes to describe a photometric band or system:

CoordSys.Band.Name	Specific name for band
CoordSys.Band.ID	URI pointing to band definition
CoordSys.Band.GenID	Generic name for band
CoordSys.Band.Description	Description of band
CoordSys.Band.PhotSysID	System that band is a member of, if any
CoordSys.Band.Zero	A submodel defining the zero point for the band and the absolute flux at the zero point.
CoordSys.Band.RefSpec	URI for reference spectrum
CoordSys.Band.EffSpCo	A submodel defining the effective spectral coordinate

The *generic* name of a band is the band name as commonly used, e.g., "U", "B", "V", etc. The *specific* name is the name of the actual filter (or other bandpass mechanism) used. *CoordSys.Band.Zero* includes specification of the value, unit, and UCD (physical type) of the measured values (typically magnitudes) at the specified zero point, as well as the flux density at the zero point.

The main point of *Band* is to identify the photometric band used for a given flux value and optionally provide sufficient information to estimate the corresponding absolute flux value. For well known photometric bands expessed in magnitudes the zero point magnitude and flux may be sufficient. In the general case more of the metadata defined in the Photometry model may be required, including detailed calibrations such as the transmission curve.

We do not attempt to go into the details of the Photometry model here; please refer to the Photometry model proposal(s) for the details [*this is an active area of discussion.*].

In a time series dataset in which multiple bands are observed it is necessary to describe each band. A separate instance of *CoordSys.Band* is required to describe each band used. How this is done depends upon the serialization. In a VOTable the GROUP mechanism would be used to group the elements of each band instance.

In the current Photometry model proposal some elements required to describe a spectral band are defined the *Characterization* data model, described in the next section.

3.1.9 Characterization Metadata

The VO *Characterization* data model is used to uniformly characterize *any* astronomical dataset in physical terms. The spatial, spectral, temporal, and sometimes polarization measurement axes are physically described in terms of coverage (location, bounds, support), and accuracy (sampling, resolution, statistical and systematic errors, level of calibration). Limited capabilities for describing the flux "axis" or observable are also provided. Additional custom characterization "axes" can also be defined.

Dataset characterization is one of the most important VO data models for data discovery. Using characterization one can look for data with a certain spatial, spectral, or time coverage, with a certain limiting resolution, etc.

The full Characterization data model is too much detail to get into here, but to give some idea of what this consists of here are some attributes for the time axis. Similar values are defined for the spatial and spectral axes.

Char.TimeAxis.Name	Name for time axis
Char.TimeAxis.Ucd	UCD for time values
Char.TimeAxis.Unit	Unit for time values
Char.TimeAxis.Coverage.Location.Value	Midpoint of exposure on MJD scale
Char.TimeAxis.Coverage.Bounds.Extent	Total exposure time
Char.TimeAxis.Coverage.Bounds.Start	Start time
Char.TimeAxis.Coverage.Bounds.Stop	Stop time
Char.TimeAxis.Accuracy.BinSize	Time bin size
Char.TimeAxis.Accuracy.StatError	Time coordinate statistical error
Char.TimeAxis.Resolution	Time resolution

3.1.10 Data Values

All of the time series metadata described in the preceding sections is used to describe available datasets in the query response. It is also used in an actual time series dataset (if VO-compliant rather than native data is returned) to describe that dataset. Actual datasets however also include the time series data values. The query response from a discovery query for time series data and an actual time series dataset can be very similar, the main difference being whether the data values are returned.

Much time series data is quite simple, e.g.

time flux flux-error

In this case the flux measurement is made in some spectral band which is specified as a constant in the metadata describing the dataset.

Also quite common is multi-band data, where flux is measured simultaneously, or nearly simultaneously, in several spectral bands or filters. It becomes necessary to identify the spectral band for each flux measurement. An additional problem is that for "nearly simultaneous" observations (where something in the instrument is physically reconfigured for each successive spectral band) we need to give the time separately for each band. The following simple representation can support all of these cases:

time1 flux flux-error band1
time1 flux flux-error band2
time1 flux flux-error band3
time2 flux flux-error band1
time2 flux flux-error band2
etc.

In this example we have three color photometry performed simultaneously for each time value. If the instrument does not permit simultaneous photomery the time will vary for each individual flux measurement but the same representation will work (if desired "close" time samples could be collected together on one row containing data for all bands for presentation to the user, even though the data is represented in the form shown).

Additional complications are possible, e.g., the time sample bin size might vary between successive samples, or a description field might be added if it is necessary to say something about the origin of each data sample, e.g., when a time series is synthesized using data from multiple sources. In general any element of the data might vary for successive data samples, in which case it must be represented as a data vector; otherwise it reduces to a constant in the dataset metadata.

In the multi-band example above, “band1”, “band2”, etc. would be band names such as “U”, “B”, “V”, etc. as defined in the *CoordSys.Band* metadata. The band metadata would more fully describe each spectrophotometric band.

The model elements (table columns) required to represent the data shown in these examples are as follows:

Data.TimeAxis.Value	Time value of data point
Data.SpectralAxis.Value	Spectral band
Data.FluxAxis.Value	Flux value
Data.FluxAxis.Accuracy.StatErrLow	Lower bound of statistical error
Data.FluxAxis.Accuracy.StatErrHigh	Upper bound of statistical error

3.2 Access Protocol

There hasn’t been time yet to flesh out this section, however the data access protocol required for time series data will be very close to what is already defined for 1D spectra (SSA).

The primary query parameters are POS,SIZE, BAND, TIME, and FORMAT, specifying the spatial, spectral, and time regions for which data is desired. Query by object name is also possible in some cases, or a prior query to a name resolver can be done client side to convert an object name into a position.

Additional standard parameters are defined to query by time resolution, variability amplitude, and target name or classification. One can also search for a dataset given its publisher or creator dataset identifier. The collection name (e.g. survey or instrument) can be used to restrict a query.

In most simple cases an entire dataset (time series in this case) is returned. However a service can provide “virtual data” capabilities to compute on the fly the output data to be returned. The simplest case of this is format conversion, e.g., returning the same dataset as a VOTable, as a FITS table, as simple CSV or TSV, rendered as a JPEG or HTML page, etc. Subsetting data is also possible, e.g., to return only a cutout in a certain time range of a very large time series. Many services will compute the output dataset to be returned on the fly, translating from whatever representation is used internally in the local archive.

An important type of virtual data generation is data model mediation. In this case a native project-specific time series is convered to the VO time series data model and this is what is returned to the client application. Although some information will generally be lost in such a conversion, it is necessary for large scale multiwavelength/multimission analysis as otherwise a client application would have to understand the native data format of every time series data collection available. Pass-through of native data can also be provided to provide greater capability for clients that can deal with the native data format of a specific data collection.

Again, this is all pretty standard and well developed at this point within the IVOA data access interfaces. Much of what is defined in the SSA interface for spectra is what will be needed for time series data as well.

4 Usage Examples

To be added: provide some examples of simple queries. (The simplest client is a Web browser; one can submit a GET query, render the VOTable query response as XML, and then go and GET the individual datasets for local analysis).

Appendix A: Data Use Cases

The idea with this section is to say something about the major time series data collections used in the design of the IVOA time series interfaces. As of this writing the major inputs considered for the work presented here included the relevant IVOA standards (Spectrum, SSA, Photometry), the SimpleTimeSeries work by Josh Bloom, the OMC and COROT archive interfaces (these are based upon an earlier version of SSA/Spectrum), and the NSted and Harvard Time Series Center facilities within the US.

References

IVOA Spectral Data Model
http://www.ivoa.net/Documents/latest/SpectrumDM.html

IVOA SSA Access Protocol
http://www.ivoa.net/Documents/latest/SSA.html

Draft Photometry data model (VAO/SAO; ESAC has developed a comparable model)
http://hea-www.harvard.edu/~jcm/vo/docs/PhotDM-20101126.pdf

Josh Bloom's SimpleTimeSeries work
http://dotastro.org/