

Extending VOEvent for more complex data

Roy Williams, Matthew Graham
California Institute of Technology

The VOEvent standard has been well taken up by the community since 2006, and there have been thousands of events in this format. Now there is interest in new schema for VOEvent 2.0, specifically to be able to put small tables, time series, ephemeris etc into the packet. We also discuss the use of linking to connect to complex data, and ask if any extra structure is actually needed, or if linking can always be the answer. We present some use cases for such extra structure, consider representing these within the current VOEvent structure, and suggest a simple table extension of VOEvent, that is already IVOA Recommendation, that would represent all the use cases in a simple way.

Introduction

VOEvents [1] are intended to communicate the discovery of a transient celestial event, with the implication that timely follow-up is being requested and reported back. It should include information that allows decisions and classifications to be made by human or machine. Since the IVOA Recommendation [1] of VOEvent 1.1 (June 2006), many astronomical projects have expressed interest in building VOEvents, and some are actually doing so, and have created hundreds or thousands of events. The next challenge is to consider improvements to the 1.1 specification, that may make the information in the packet more useful, more clear, or more comprehensive; but without putting an undue burden on the parsing software or invalidating existing version 1.1 events. In this note, we consider some use cases for VOEvent and the data structures that would be appropriate.

One of the objectives of VOEvent is to establish it as an alternative way to publish first discovery, in the way that Astronomer's Telegram [2] and CBAT [3] are. These event notices are usually English text, unreadable to machines, but some have a small table or list as part of the notice. If we believe that VOEvent should be able to express whatever these other publications can express, then VOEvent should be able to hold this extra data, and furthermore, we should make that data structured, and thus accessible to machines also.

VOEvent is an XML model. Other serialization possibilities are also in use, such as JSON and RSS/Atom feeds, these last being XML dialects that are not compatible with VOEvent 1.1. However, the point of this note is not to question or suggest syntax for VOEvent, but to ask if more complex data extensions should be in it. If the answer is yes, these extensions can also be represented – to some extent – in these other kinds of syntax.

The VOEvent data model from 2006 has several components: `<who>`, `<what>`, `<wherewhen>`, `<How>`, `<why>`, together with `<Citation>`, `<Reference>`, and `<Description>`. The `<Reference>` element is to contain links to documents and data objects that give further information about the event. The `<why>` section assigns one or more causes of the event, in a probabilistic way, to the possible astrophysical processes, where the names of those processes can include concepts from an IVOA-controlled standard vocabulary.

Most of these do not concern us in this note, but specifically:

`<Reference>`

VOEvent defines a `<Reference>` element, which allows linking from the event packet to external resources by the URI mechanism, for example:

```
<Reference uri="http://gcn.gsfc.nasa.gov/gcn/notices/sw0041000msni0284.fits" type="url"/>
```

This allows web links, so that a complex data object (eg FITS file) can be available to readers of the VOEvent, but without it being carried everywhere and stored in event repositories. Currently the values of the "type" attribute are a closed enumeration: `voevent`, `url`, `rtml`, and `ivorn`. In VOEvent2, this is expanded to an open form, properly specifying the *semantic* nature of the linked resource. For example:

```
<Reference
  type="http://ivoa.net/rdf/formats#fits-image"
  uri="http://.../sw0041000msni0284.fits"/>
```

identifies by the *type* attribute that the reference is to a FITS image. A private vocabulary could also be used:

```
<Reference
  type="http://my.home.page/definitions#36"
  uri="http://my.home.page/images/7463736252.jpg"/>
```

to say that this is a Johnson V-band image of a globular cluster -- since that is what 'definition#36' resolves to. The typing of the link can always be resolved if you are unfamiliar with it to a human-readable description. This means that parameters and hypotheses in the packet can be defined in the formal terms of the Semantic Web, rather than just natural language descriptions of what the data means.

In VOEvent, it is always recommended that big data be linked rather than included in the VOEvent. Small data objects can also be linked in this way. However, in many cases (eg a 3x3 table), there may be reasons to include it directly with the packet: perhaps the event author does not wish to set up a web server and guarantee permanent links, and perhaps the intended recipients cannot fetch URL links. Therefore, in this paper, we are considering a way to put structured data in the VOEvent packet that are *small*, in the above sense.

<What>

The `<what>` element is to contain data from a vocabulary that is specific to a given event, or type of event. In the VOEvent model, parameters are defined in the `<what>` section through a text description, UCD, unit, etc so that event consumers can know precisely what is meant, for example

```
<Param name="Sun_Distance" value="132.21"
  unit="deg" ucd="pos.angDistance">
<Description>Angular distance from the sun to the event position</Description></Param>
```

Thus the Parameters, and a mechanism for grouping them, can be used by an author to define precisely the data they wish to express. Parameters can also be collected into a `<Group>` element, to make relationships clear. For example a quantity and the error in that quantity can be Grouped together to show their close connection.

Learning from the past: the FITS format

In moving forward with VOEvent, we may be able to take lessons from the FITS standard. While there are still matters of concern, FITS has been remarkably durable and effective, with essentially all working astronomers having some knowledge of it. FITS is therefore a model to be imitated.

The metadata object of a FITS file is a keyword-value pair, similar to the `<Param>` construct in VOEvent, except that the FITS metadata describes a multi-dimensional array, and the `<Param>` describes an event. There is no schema or complex data model inherent in the way that FITS metadata is kept, nothing but names and values. Yet structure and standards emerge when a community uses this clay to build data and tools. The keywords `BITPIX`, `NAXIS`, `NAXIS1`, `NAXIS2`, etc are in all FITS files to explain how to read the data array, just as VOEvent has sky position and time as requirements for all events. Some conventions have become very widespread, for example the WCS keywords `CTYPE1`, `CRVAL1`, `CD1_1`, etc that define how the data array maps to the sky.

However, a FITS file is not fully machine-readable, in the semantic sense. There is no widely-accepted way to find the meaning of the keywords: the file says `FWDX3=253`, but the term is not defined. Some

attempts have been made to build a formal dictionary, for example [4], but nothing is widespread. In VOEvent, we intend to correct this problem, as described in the next section.

Event Stream

In order to control the collection of Parameters used in the `<what>` section, the concept of “Stream” has been proposed [5], which separates the definition of the Parameters from their values. This provides for authors to build a registry of the keywords (i.e. Param and Group names) that will be used in their events, with machine-readable descriptions of each. This we call an event 'Stream', which will be available through the IVOA registry. So this is the bargain: in exchange for a universal way to find the meaning of all the keywords, all the keywords must be defined in the Stream, and event providers cannot simply use new keywords without updating the Stream record.

The idea is that all VOEvents that come from a specific instrument or project will all have the same structure, so that parameters and their meaning are defined in advance, and kept in the VO Registry for reference. Then when an event is actually issued, only the parameter name, group, and value are needed.

However, Parameters are intended only for primitive data objects (float, integer, string etc), and we might consider an additional more complex, structured object as part of VOEvent 2.0. In this note, we collect use cases from different projects to ask if the current and proposed VOEvent structure can carry the kind of data objects that the authors wish to express. The intention is to use these use cases as a way to understand what extensions might be added to the VOEvent 2.0 specification. We take several examples from published reports in the Astronomer’s Telegram, also called ATEL [14].

Use cases for complex data

- Light curve

Crucial information about a celestial object is contained in its light curve, which is a history of its (photon) output. Machines can extract the light curve and make decisions on the nature of the source. For example, from ATEL 2323, we find this light curve:

Date (start/stop)	Exposure [ks]	Ph.Index	F(2-10 keV) [erg/cm ² /s]
2009-12-01 16:38-16:56	1.0	1.58+/-0.05	1.14e-10
2009-12-02 00:30-00:50	1.2	1.56+/-0.05	1.32e-10
2009-12-03 15:25-15:40	1.0	1.60+/-0.06	1.22e-10
2009-12-04 00:43-02:43	2.7	1.52+/-0.03	1.58e-10

Another light-curve (ATEL 2289) shows information in addition to optical magnitude, the distance, phase, number of observations, and names of observers.

UT Date	r(AU)	d(AU)	Phase	R_mag	Exp.	Observers
Oct 10	1.090	0.289	64.2	15.37	34	Barajas
Oct 11	1.085	0.276	64.6	15.28	38	Foster
Oct 17	1.056	0.202	67.7	14.79	31	Somers, Hicks
Oct 22	1.031	0.150	72.2	14.10	53	McAuley, Shitanishi

There are many examples of small light curves that can be found in the Astronomers Telegram or CBAT notices, so this is an important use case.

- Object list

We consider some requirements forming within the LSST group [6], where some suggestions (L. Jones) include this:

- RA/Dec \pm dRA/dDec of diasource
- Mag \pm dMag of Diasource
- LSST catalog objects (stars/galaxies) within X " of this RA/Dec (median mag in ugrizy and rms mags or other variability info in catalog, plus their RA/Decs, plus their star/galaxy likelihoods)

- other diasources within Y'' of this RA/Dec in the past Z number of days (their Ra/Dec/mags)

Here we see the suggestion that the event report should contain a list of nearby objects in the sky, and another list of observations of the objects nearby in time. Other examples of object lists could come from listing known nearby galaxies or radio sources, as these can change the priors for the classification process.

- Frequency series

The LOFAR [7] radio antenna array will be coming to full operation in 2010, and the team, in preparation for routine publication of radio transients, has defined an event template (an event “stream”) for those transients. What they wish to publish is a set of radio frequency bands, each defined by a minimum and maximum frequency, and flux density. While the current stream has just two bands (Band1 and Band2), there is a desire to have multiple bands. This could be described as a table, with columns: min frequency, max frequency, and flux, and a row of the table for each band.

- Ephemeris for a minor planet

A major driver for research in transient astronomy is the discovery of Near-Earth asteroids. After discovery, an ephemeris is published, being the future position of the asteroid as a function of time. The Minor Planet Center has a service to compute such ephemerides, with results like this:

Year	day	h	min	---RA---	---Dec---	elong	mag	RA(sec/min)	Dec(sec/min)
2010	03	17	12	09 22 38.3	+44 17 19	126.4	20.6	-0.19	-2.60
2010	03	17	16	09 22 35.4	+44 06 55	126.4	20.6	-0.18	-2.60
2010	03	17	20	09 22 32.6	+43 56 30	126.3	20.6	-0.17	-2.61
2010	03	18	00	09 22 29.9	+43 46 04	126.3	20.6	-0.16	-2.61
2010	03	18	04	09 22 27.4	+43 35 38	126.3	20.6	-0.16	-2.61

The detection and follow up of such unknown asteroids is a major driver for funding transient surveys, and should therefore be expressible with VOEvent. Such a time series does not contain observations, only predictions; so we would not expect, for example, anything about the observational equipment or filters used, since no observations have been taken.

- Classifications for asteroids

In ATEL 2372, there is a set of possible classifications for a newly discovered asteroid, based on comparison between spectra. The asteroid is classified according to its closest spectral match among known asteroids:

MISFIT	OBJECT NAME	TAXONOMIC CLASS	
		(THOLEN)	(BUS)
0.61	93 Minerva	CU	C
0.97	3192 A'Hearn		C
1.10	1017 Jacqueline		C
1.13	1041 Asta		C
1.18	5294 Onnetoh		X

- Probability density from gravitational wave detector

The LIGO project [8] is building software pipelines to create alert events for the broad community, in case of a discovery of a burst or inspiral source. Given that the detection comes from four instruments with different locations, orientation, etc, it means that the sky localization is not compact, but rather is a probability map over a wide area [9]. The LIGO team is already developing a representation of this probability density in terms of a set of “tiles”, each 0.5 degrees square, with an associated probability. Thus a LIGO event will include this table of tiles, from which follow-up observers will derive an observing plan.

Representing the use cases

With existing VOEvent structure

In this section, we will describe how each of the use cases could be met within the existing VOEvent (v1.1) data model. We also consider the “stream” concept (as represented by the VOEvent registry infrastructure [5]) to be part of the existing data model.

- Light curve

Each column in the light curve can correspond to a Param element in the `<what>` section and each row in the table to a Group of these Params: for example, the first row becomes an object called “row1”:

```
<Group name='row1'>
  <Param name='Date' value='2009-12-01 16:38-16:56' />
  <Param name='Exposure' value='1.0' unit='ks' />
  <Param name='Ph.Index' value='1.58+/-0.05' unit='' />
  <Param name='F(2-10 keV)' value='2009-12-01 16:38-16:56' unit='erg/cm2/s' />
</Group>
```

Unfortunately this means that all the rows have their own name, and it breaks the Stream model, since each possible Group name would need to be defined there, and they are potentially unbounded: row1, row2, row3,

- Object list

Two types of list are specified in the use case: a list of nearby objects in space (within 30 arcsec.), and a list of the objects nearby in time (observations of objects within the last 24 hours). The first type can be addressed using the same approach described in section 3.1.1 by mapping each object to a Group of Params, one for each object attribute, although, as noted above, it would not be possible to build a Stream record for that structure.

- Frequency series

The table can be represented as in section 3.1.1 with each band (row) corresponding to a Group containing Params for min frequency, max frequency and flux, although, as noted above, it would not be possible to build a Stream record for that structure, unless the number of frequency bands to be used is locked in at the Stream level.

- Ephemeris

Philosophically, the correct section for this type of information is the `<wherewhen>` section since it relates to targeting in spacetime. Rots [10] describes how this can be achieved within the framework of STC (Space Time Coordinates, [11]). See also [12], for an example of an ephemeris representation that uses over 300 lines of XML for three lines of ephemeris.

Classification for asteroids

Philosophically, the correct section for this type of information is the `<why>` section since it deals with scientific characterization, in this case, the set of possible classifications for a newly discovered asteroid. Each possible classification would correspond to a separate `<Inference>` element. One problem is that the only way to express a quantitative measure of the plausibility of a particular `<Inference>` is through its “probability” attribute, which takes a floating-point value between 0.0 and 1.0 (inclusive). In the use case, however, the plausibility of each classification is specified through the value of the “misfit” parameter, which is some goodness of fit measure. These values could be represented in the `<what>` section as a set of Params but this is not likely to be a list of known cardinality and so, as above, it cannot be described in the stream specification.

- Probability density map

As noted above, the LIGO team have decided to represent the density sky map as a collection of tiles, perhaps 0.5 degree square, each with probability and other information. Each tile could then be expressed as a Param or Group of Params in the `<what>` section, although (as above) this cannot be expressed in the stream definition because of the potential unknown cardinality of the list. The more natural location for this information is in the `<wherewhen>` section and although STC can represent footprints, it does not associate probabilities with these, so it would be difficult this way.

By extending VOEvent 1.1

VOEvent is intended to be both structured and pragmatic and, although it is possible to meet many of these use cases with the existing data model, the resulting representation tends to be rather unwieldy: for example, a Group of Params for each row in a table. The open-ended nature of (tabulated) lists is also somewhat at odds with the required cardinality imposed by the stream model and so limits the amount of validation that the stream concept facilitates. Finally, there are areas where the current VOEvent data model does not meet the requirements of the use case and an extension is necessary.

It should be noted that one alternative solution to representing the data indicated by all these use cases is to simply include the text, much as it is represented above. This would allow humans to see the data, in much the same way as they do now when reading an ATEL or CBAT notice; but it does not give machines meaningful access.

An existing schema for tabular data

All the use cases in section 2 can be met by extending the VOEvent 1.1 data model to include a simple representation for tabular data. VOTable [13] is the default IVOA standard for representing tabular data but is perhaps too complex for these purposes (certainly not all features of it are required here) and is known to have a history of validation and code-binding issues. For these reasons, other IVOA WGs have defined simple tabular representations, e.g. TableSchema in the VODataService standard from the Registry WG.

A very simple table schema is available as part of STC [11] . Another advantage of using the STC is that existing rigorous standards for space-time coordinates, already used in VOEvent, can also be used for subclasses of the table, such as light curves and other time series.

```
<STCTable xsi:schemaLocation="stcTab.xsd
  http://hea-www.harvard.edu/~arots/nvometa/v1.30/stcTab.xsd">
<Table>
  <!-- Header definition -->
  <Header>
    <TH id="Time">Time</TH>
    <TH id="FDVal">FluxDensity</TH>
    <TH id="FDErr">FD Error</TH>
  </Header>

  <!-- and table body -->
  <TR><TD>123.4</TD><TD>56.7</TD><TD>.98</TD></TR>
  <TR><TD>133.4</TD><TD>46.7</TD><TD>.98</TD></TR>
  <TR><TD>143.4</TD><TD>36.7</TD><TD>.98</TD></TR>
  <TR><TD>153.4</TD><TD>26.7</TD><TD>.98</TD></TR>
  <TR><TD>163.4</TD><TD>16.7</TD><TD>.98</TD></TR>
</Table>
</STCTable>
```

Conclusions

Use cases from across the community provide a requirement to represent tabular data/information within a VOEvent packet. Although some of these can be met within the framework of the existing VOEvent data model using deconstructed representations, these are not largely pragmatic. There also

remain a number of use cases which cannot be met in this manner. Therefore it seems reasonable to propose extending the VOEvent 1.1 data model in the following manner:

- Introduce a simple tabular representation based on the existing tabular construct within the IVOA STC specification as another allowed element within the `<what>` section.
- Open the "type" attribute of the `<Reference>` element to take a URI value. This allows any reference to an external resource to be semantically typed and so arbitrarily (large) complex data constructs can be linked to in a meaningful manner.
- Add an additional attribute to the `<Inference>` element in the `<why>` section to allow a more general plausibility measure to denote goodness of fit, etc.

References

- [1] <http://www.ivoa.net/Documents/latest/VOEvent.html>
- [2] Astronomer's Telegram, <http://astronomerstelegram.org>
- [3] Central Bureau for Astronomical Telegrams, <http://www.cfa.harvard.edu/iau/cbat.html>
- [4] <http://iraf.noao.edu/projects/ccdmosaic/imagedef/fitsdic.html>
- [5] <http://www.ivoa.net/internal/IVOA/IvoaVOEvent/VOEventStream-0.2.html>
- [6] <http://www.lsst.org/>
- [7] <http://www.lofar.org>
- [8] <http://www.ligo.org>
- [9] A.C.Searle, P.J.Sutton, M.Tinto, G.Woan *Robust Bayesian detection of unmodelled bursts*, <http://arxiv.org/pdf/0712.0196>
- [10] Rots, A., 2009, *XML for MPC Ephemerides and other Time Series*, <http://www.cacr.caltech.edu/hotwired2/program/presentations/MPCTimeSeriesRots.pdf>
- [11] <http://www.ivoa.net/Documents/latest/STC.html>
- [12] <http://hea-www.cfa.harvard.edu/~arots/nvometastc/MPC.xml>
- [13] <http://www.ivoa.net/Documents/VOTable/>