

Vocabularies and Semantics for VOEvent

Frederic V. Hessman

Institut für Astrophysik, Georg-August-Universität Göttingen

Norman Gray

Department of Physics and Astronomy, University of Glasgow

This book is about real-life science and practical implementations of strategies for catching and processing astronomical events. The gentle reader is probably asking herself “why in the world is there a section about something as obscure and irrelevant as ‘semantics?’” The answer is actually quite simple: the “semantic” part of the problem is concerned with making explicit the meaning of a VOEvent message, because the whole point of VOEvent is to transport meaningful content. Wiktionary [1] defines semantics as “the study of the relationship between words and their meanings” or (to paraphrase slightly) “the individual meanings of words, as opposed to the overall meaning of a message.” Thus, semantics in VOEvent is 1) the attempt to clearly identify the astronomical descriptions used to characterize an event, and 2) to put those meanings in a clear and machine processable context relative to the other content of the message and the scientific point of dealing with such messages. Computers are very good at processing strings and transporting bits – semantic technologies are about making meanings just as conveniently processable.

Semantics has become important in VOEvent and modern astronomy in general for a very simple reason: as long as astronomical information was being processed by humans, one could count on there being both senders and receivers with the right astronomical training capable of insuring that the content was reasonable and intelligible, and either phone or email access to ask “what's column X supposed to mean?”. In the age of massive astronomical surveys, this is no longer possible. While astronomers can count on it being possible to transfer terabytes of images, spectra and tables nearly seamlessly between themselves and their applications, there is no really universal means of saying that an image is of a galaxy, the spectrum is of a brown dwarf, and that the table contains the wavelengths of telluric lines. Both technical and social changes mean that data is now shared more promiscuously, and users and their computers receive it from multiple sources, so that informal semantic arrangements (the “what is column X?”) are becoming inadequate. The whole point of VOEvent is to permit computers to produce and process astronomical events, but since many producers are capable of emitting reports on a wide range of astronomical events and most receivers are interested in only a small subset of all possible events, being able to clearly describe and recognize the semantic content of a VOEvent is crucial to making the protocol generically useful.

There are several different forms and levels of the organization of semantic information. The ultimate semantic solution is to teach computers to process information in some semblance to how an astronomer would do it, but pending some unbelievable development in artificial intelligence, this is unlikely to happen soon. In any case, such an elaborate machine ‘understanding’ is not necessary. A real astronomer knows what a “spiral galaxy” is, that a real instance has a position on the sky, angular and geometric size, brightness and other properties. A real astronomer can guess that “SN Ia” probably means that this is a label for a Type Ia supernova and that they occur in spiral galaxies. These fairly lightweight structural relations are what we hope to encode at this initial stage; more sophisticated ‘understanding’ -- such as the knowledge that an object which possesses an RA and a Dec must be an astronomical object of some type -- is perfectly feasible, and usefully deployed in other disciplines such as biology, but more than we need to worry about right here.

We can distinguish various levels of formality here (and the descriptions are formalised in, for example, [2] and the more compendious [3]. A ‘controlled vocabulary’ is a simple list of approved terms and descriptions; a ‘taxonomy’ and a ‘thesaurus’ have more or less additional hierarchical structure, such as broader/narrower relations (‘compact object’ is a broader term than ‘quasar’) or relations (such as the famous Yellow Pages cross reference, ‘Boring; see Engineers’). The standard definition of an ‘ontology’

is “an explicit specification of a shared conceptualization” (after [4]): this is a useful definition, but it covers a huge range of possibilities, all the way down to controlled vocabularies, and so it is often restricted to the case where a thesaurus has a ‘formal IsA’ relation. The distinction is simple: saying, in an ontology, that compact objects have a subclass called quasars means that anything which is a quasar is also a compact object, as a matter of logical necessity; saying, in a thesaurus, that ‘cars’ have a narrower, or more specialized, concept of ‘wheels’ does not imply that everything that is a wheel is also a car! Loosely, thesauri are useful for searching, and ontologies are necessary for making deductions. Most, but not all, of the examples we mention below are thesauri, but for our purposes we will lose no precision in referring to them indiscriminately as ‘vocabularies’.

There are several examples of astronomical taxonomies in common use: the NASA Taxonomy [5] keeps an official list of terms used to describe a wide variety of information from various missions and organizations to “audiences” and “work breakdown structures”; and the “Astronomy Visualization Metadata” (AVM) taxonomy [6] is an international standard used to describe the content of astronomical outreach media (permitting an image of a spiral galaxy to be labelled as such). The latter has a very clear hierarchical structure. Also, the IAU produced an official thesaurus of astronomical terms in 1993 (<http://www.aao.gov.au/lib/thesaurus.html>) for the benefit of librarians, containing not just a list of words but their translation into several (European) languages and a modest level of connections between terms (“narrower” and “broader”). Finally, there is an Ontology of Astronomical Object Types [7] which has been produced under the auspices of the IVOA—this is intended for reasoning rather than search, but illustrates the intricacy of what is possible.

Historically, the semantic content of a VOEvent package was based on the assumption that the producers were specialists for one type of event—e.g. gamma-ray bursts or supernovae—and that the receivers knew their producers and could expect to process implicitly defined and well-known content. The only real semantic content that was clearly defined was the simple “Unified Content Descriptors” (UCD) system created by the Virtual Observatory to identify the rough content of tabular data - things like “em.wl” for “wavelength”, “pos.eq.ra” for “Right Ascension”, “phot.mag;em.opt.R” for R-magnitude, “meta.table;meta.id” as the closest description for “catalogue number”. This is important semantic content supported in the <What> section of a VOEvent via the <Param> entries, each of which can be assigned one or more UCD values. However, this is semantic content of the most primitive kind describing the content at the lowest level of a VOEvent. In the <Why> section of a VOEvent, the <Concept> entries were provided for transporting higher-level semantic content, e.g.

```
<why probability="100">
  <Concept type="name">Tycho's Stella Nova</Concept>
  <Concept>SN Ia</Concept>
</why>
```

but the content itself was not based on any clear system: if one saw "SN Ia" in one message and "Type Ia supernova" in another, it was left to the poor receiver, or their regular expressions, to be clever enough to catch all the expected and unexpected possibilities, or risk not catching an interesting event.

One might have expected that, since all of these semantic problems are common to all Virtual Observatory projects and not to VOEvent alone, the IVOA would have long since solved these problems by creating clear semantic standards. To some extent, this is true: the IVOA has formalised UCDs as a vocabulary for astronomical data (note that UCDs, deriving ultimately from a census of column names at CDS, form a taxonomy of astronomical data, not of astronomical objects -- thus there is a UCD for an ‘abstract’, but no UCD for a ‘star’), and has adopted SKOS (see below), but there are still important parts of the jigsaw missing. The much-heralded “utypes” will provide a syntactic bridge for linking data and concepts, but the meaning of “data model” and the means of publishing what such things are supposed to be -- true semantic content or mere software class descriptors -- is different depending upon whom one talks to. Thus, the always practically minded VOEventist is yet again faced with the problem of a well-defined task with an apparently non-standardized solution path. The IVOA's adoption of SKOS, as a means of describing and relating vocabularies, is an important step on the route out of this thicket, and we describe it in some detail below.

Note that here and below we are careful to talk of vocabularies, plural. Though it may seem obviously useful to have a single master vocabulary, such a thing is unwieldy, would remain unfamiliar to most scientists, and would require a major effort to agree on. Instead it is reasonable to have multiple smaller, and more specialized, vocabularies, which can be developed in a more agile fashion, and which can be interrelated using standardized SKOS structures.

This section treats two fundamentally semantic and yet very practical aspects of VOEvent: namely the creation and the recognition of intelligible semantic content in an event message. First, we identify where the semantic information is contained within the elements of a VOEvent document. Then, a simple means of publishing the possible content is described. Finally, the practical aspects of using the published content to produce and process VOEvent documents is outlined. While some of the details may only apply to early versions of VOEvent, the general ideas should remain valid for the foreseeable future.

Vocabularies in the VO

While the IVOA has not defined what vocabularies could be used within a semantic context like VOEvent, it has specified how such information should be published. Following the work being done for the “semantic web”—developing the structures to communicate meaning on the web—the IVOA has recommended using the W3C’s “Simple Knowledge Organization System” or SKOS. The W3C describes SKOS thus: “Using SKOS, concepts can be identified using URIs, labeled with lexical strings in one or more natural languages, assigned notations (lexical codes), documented with various types of note, linked to other concepts and organized into informal hierarchies and association networks, aggregated into concept schemes, grouped into labeled and/or ordered collections, and mapped to concepts in other schemes” [8]). In short, SKOS provides a standardized framework for sharing and documenting vocabulary lists. This is obviously exactly what we need for the IVOA in general and VOEvent in particular, so the vocabulary document produced by the IVOA semantics workgroup mainly described how SKOS is used and gave a few examples.

This sounds fairly simple and, in principle, it is. Unfortunately, vocabularies published using the SKOS standard can be expressed in different formats and the IVOA didn’t place any restrictions on or make any recommendations for the formats of choice. In practice, astronomical vocabularies are likely to be published/accessed in one of two formats: RDF/XML and “Turtle”. NASA, for instance, publishes its SKOS taxonomies in RDF/XML (“RDF” means “Resource Description Framework”, and RDF/XML is a serialization of RDF into something which can be read by an XML parser). Here is a section from the NASA “locations” taxonomy [9], which illustrates how a vocabulary can be published in a very real-world semantic context:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
  xmlns:skos='http://www.w3.org/2004/02/skos/core#'
  xmlns:nt2='http://nasataxonomy.jpl.nasa.gov/cvFields#'
  xmlns:dcterms='http://purl.org/dc/terms/'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
  <skos:Concept rdf:about='loc:83'>
    <skos:prefLabel>Phoebe</skos:prefLabel>
    <skos:broader rdf:resource='loc:65' />
    <nt2:status>Approved</nt2:status>
    <nt2:type>Descriptor</nt2:type>
    <nt2:code>83</nt2:code>
    <nt2:inputdate>2004-05-01</nt2:inputdate>
    <dcterms:dateAccepted>2004-06-11</dcterms:dateAccepted>
    <dcterms:modified>2004-06-11</dcterms:modified>
  </skos:Concept>
  ...
</rdf:RDF>
```

Anyone familiar with XML will recognize the content right away, even without a previous knowledge of SKOS. The first line simply states that the document is XML. The <rdf:RDF> element has attributes describing the namespaces used, e.g. the RDF/XML and SKOS definitions as well as the standard Dublin documentation standards. The rest is a single SKOS concept for a location with the cryptic label “loc:83” (this is a URI, though unfortunately the namespace 'loc:' has not been declared in the RDF file -- we can guess it should be something like:

```
xmlns:loc='http://nasataxonomy.jpl.nasa.gov/xml/locations.skos#')
```

Each concept has one or more preferred labels <skos:prefLabel> that are human-readable strings. In this case, the preferred label tells us that location loc:83 is actually an astronomical object called “Phoebe” (FOOTNOTE: more properly, loc:83 names “the concept of Phoebe”: Phoebe is a thing you find in the sky; the concept of Phoebe is something you find in a library index or on a computer). Solar system experts will immediately recognized that Phoebe is one of Saturn's moons and hence of obvious interest to NASA. This concept has a single ontological entry in the form of a “broader than” reference to the concept “loc:65”, which -- unsurprisingly -- is Saturn. The other concept entries are NASA internal things (<nt2:...>) or purely documentation items expressed using the internationally accepted Dublin Core vocabulary (<dcterms:...>). One immediately notices that the real astronomical content of the <skos:Concept> boils down to a very small number of things:

- the name “loc:83” is associated with the astronomical object “Phoebe”
- the hierarchically broader concept associated with “loc:83” is “loc:65”, “Saturn”

The rest is just syntax information and documentation. Note that “loc:65” is the concept for Saturn, not the string “Saturn”. This is good for several reasons: “Saturn” can mean other things in other semantic contexts; “Saturn” might be spelled differently in different languages (NASA's loc:65 Concept could have included the labels <skos:prefLabel xml:lang='en'>Saturn</skos:prefLabel> and <skos:prefLabel xml:lang='zh'>土星</skos:prefLabel> appropriate for Chinese end-users). The label “loc:83” thus remains an unambiguous and universal label for a clear astronomical concept, one of Saturn's moons. To what extent that Saturn can be considered a “broader” term than one of its satellites is defined by the context: in somebody else's taxonomy, the broader term might have been “moons of planets”, but the creators of the NASA taxonomy were very happy to have an official list of locations with a minimum of ontological baggage.

The same document expressed in “Turtle” looks like the following:

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix nt2: <http://nasataxonomy.jpl.nasa.gov/cvFields#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
<loc:83>
  nt2:code "83";
  nt2:inputdate "2004-05-01";
  nt2:status "Approved";
  nt2:type "Descriptor";
  dcterms:dateAccepted "2004-06-11";
  dcterms:modified "2004-06-11";
  a skos:Concept;
  skos:broader <loc:65>;
  skos:prefLabel "Phoebe".
```

You may or may not find this more readable; it clearly requires a specialised parser. What is possibly not obvious is that the RDF/XML form also, strictly, needs a special parser. The RDF/XML format is a flexible one, and for example permits the string-valued relations to be expressed as attributes:

```
<skos:Concept rdf:about='loc:83'
  skos:prefLabel='Phoebe'
  nt2:status='Approved'
  nt2:type='Descriptor'
```

```
nt2:code='83'  
nt2:inputdate='2004-05-01'  
dcterms:dateAccepted='2004-06-11'  
dcterms:modified='2004-06-11'  
<skos:broader rdf:resource='loc:65' />  
</skos:Concept>
```

Few RDF/XML writers take advantage of the flexibility the format offers, and so these files can often be processed as 'just XML', but one should be aware that doing so is not completely reliable.

NASA's RDF/XML "taxonomies" stand alone in the sense that they define a semantic context with well-defined entries but which make no references to a higher semantic level. For instance, "loc:3" is "Planetary Systems" around the Sun, but there is no way to connect this concept to exoplanetary systems (before the Kepler mission, this wasn't a problem for NASA scientists, but they will presumably have to revise their location vocabulary). This shows that scientific vocabularies are dynamic objects whose versions must be updated and maintained. The IAU 1993 thesaurus suffers considerably from signs of aging; the IVOA is attempting to rectify this by creating and maintaining a new thesaurus (LINK!).

The other obvious observation is that the NASA taxonomies probably do not correspond to ESA's location taxonomy (not published, if it exists). Two different organizations are likely to create different vocabularies for the same purpose but with very different formats, ranges, and limits. When NASA and ESA computers want to cooperate, there must be a means by which ESA can understand that NASA means "Saturn's moon, Phoebe" when it says "loc:83" and vice versa. Thus, it is generally not enough to simply publish your vocabulary, you should also aim to publish a mapping of (at least part of) your vocabulary into another one. Until the IVOA publishes its own all-purpose master vocabulary, the only relevant master translations are to the AVM taxonomy and the IAU thesaurus of 1993, despite the limitations and errors of the latter, as published by the IVOA semantics working group.

Mapping is a long-standing problem: you can either map into and out of a large vocabulary, or you can publish a set of peer-to-peer mappings. There is no simple answer to which is best.

Even when large vocabularies like NASA's taxonomies or the IAU thesaurus exist, this does not mean that one should not publish and use one's own: master vocabularies are, by definite, monstrous things with many useless entries for any particular purpose (VOEvent probably doesn't need a special vocabulary entry for the concept of an armillary sphere, present in the IAU 1993 thesaurus). A special-purpose vocabulary is easy to define, easy to publish, easy to use, but can still be connected to the larger astronomical world outside via references to external vocabularies. Managing the trade-offs here is an engineering decision, not an exotic semantic one.

The practical steps in creating and using vocabularies within the VOEvent context are described in the remaining subsections.

Step 0: Where is the semantic content of a VOEvent document?

The VOEvent Version 1.1 protocol contains several parts whose content can and should be placed in a clear semantic context. "Clear" here means that the recipient of a VOEvent document has the possibility of identifying the terms used by the creator in a manner suitable for automatic processing by a computer and with a minimum of human pre-intervention or guesswork.

<What>

This section is obviously the heart of a VOEvent document, since it describes in technical detail what the event consists of in detail.

The <Param> elements used within <What> are the fundamental units of information and have the attributes "name" (string), "ucd", "value" and "unit". While the UCD values should be taken from the official IVOA UCD list (LINK!) only, the values are presumably some real numbers, and the units are supposed to be taken from the Vizier list and grammar of standard units, it is obviously beneficial if the

names of the parameters can be uniquely identified. Some parameters will be highly specific to some instrument, others like fluxes and epochs are often easily described by the UCD, but others may be standard concepts for which a random name might hide the semantic content. For example, “ucd=phot.mag;em.opt.R” says that the parameter describes the brightness of something in an R-filter, but not whether that filter is a Johnson, Bessel, Sloan, “R” from RGB, or some other reddish filter. Formally, the VOEvent protocol lets the publisher communicate such details in the form of a <How> element, e.g. via a link to an RTML document (LINK!), but in practice VOEvents should be as self-containing as possible.

The <Group> element is used with <What> to group <Param>'s into meaningful units and so have the attributes “name” and “type”. The VOEvent standard does not place any restrictions on either attribute, but a publishable semantic description of such is again obviously advantageous. With a single <Param> value, it may be possible to guess the semantic content by the UCD term only (as with the red magnitude above), but for <Group>s this is unlikely to occur for if there is no standardized way of expressing the content of the “type” attribute.

<Why>

This VOEvent element obviously contains the most semantic information: this is the whole point of this element, but also the reason that its semantic usage has been minimal within the burgeoning VOEvent community. A <Why> element contains one or more <Name>'s, <Concept>'s, and <Inference>'s, fields.

The VOEvent standard says “the value of a <Concept> element uses a controlled vocabulary to express the hypothesized astrophysics” but admits that such a controlled vocabulary does not yet exist. In fact, the goal is to create a VOEvent-specific controlled vocabulary for VOEvent version 2.0 definition, in which case the possible form of the <Concept> content will be clearly defined.

The <Name> element is generally to be used for object names, e.g. those recognized by SIMBAD or NED, so semantic contextual information should not be placed here but in a <Concept>.

The <Inference> element has the optional attributes “probability” and “relation”. The former is clearly a number (between 0 and 1) but the latter “... is a natural language string that expresses the degree of connection between the <Inference> and the event described by the packet” and so contains obvious semantic content which needs to be clearly specified. The VOEvent 1.1 document suggests using things like “associated” and “identified” but places no normative constraints on the field, making it difficult to publish and interpret. Standardisation would be useful here: there is an extensive literature on 'argumentation ontologies', which covers more than would be required here, but which could be cherry-picked usefully.

Thus, the semantic content of a VOEvent document can be reduced to

- the name attributes of <Params> (in <What>, only partially covered in a standard VOEvent vocabulary)
- the name and type attributes of <Groups> (Ibid.)
- the content of <Concept> (in <Why>, easily covered with vocabulary)
- the relation attribute of <Inference> (in <Why>, best covered with a standard VOEvent vocabulary)

In order to understand how this content can be placed in a simple and computer-manipulatable semantic context, one first has to get to know the means by which the semantic content can be documented in general.

Step 1: Choose/Publish your vocabulary

The simplest choice of vocabulary is one already defined and published and appropriate to your needs. Within VOEvent, you have the choice of the IAU thesaurus and the AVM taxonomy published as

example vocabularies by the IVOA semantics working group, the vocabulary used by the Catalina Sky Survey, and -- hopefully available by the time this book is published -- the prototypical VOEvent master vocabulary.

If you need a different vocabulary, you will have to publish your own. Using the NASA taxonomies as examples, this might be something like the following: imagine that you operate a visual sky survey project called "Blue & Red" only capable of distinguishing between blue and red variable stars. You choose to define your own vocabulary for the three types of variable objects you can find -- blue, red, and white stars -- and for the two types of measurements made, the red and blue visual magnitudes:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF
  xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
  xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
  xmlns:skos='http://www.w3.org/2004/02/skos/core#'
  xmlns:dcterms='http://purl.org/dc/terms/'
  xmlns:foaf='http://xmlns.com/foaf/0.1/'
  xmlns:iau93='http://www.ivoa.net/rdf/Vocabularies/IAUT93#'
  xml:base='http://www.blueandred.org/rdf/BlueAndRed/v9.3'>
  <skos:ConceptScheme rdf:about="">
    <dcterms:title xml:lang="en">BlueAndRed</dcterms:title>
    <dcterms:description xml:lang="en">
      A vocabulary for the "Blue & Red" variable star survey.
    </dcterms:description>
    <dcterms:creator>
      <rdf:Description><foaf:name>Johanna Kepler</foaf:name></rdf:Description>
    </dcterms:creator>
    <dcterms:created>2010-12-25</dcterms:created>
  </skos:ConceptScheme>
  <skos:Concept rdf:about='#redVars'>
    <skos:prefLabel xml:lang='en'>red variables</skos:prefLabel>
    <skos:definition>Variables that are brighter in the red than in the
blue.</skos:definition>
    <skos:broader rdf:resource="iau93:VariableStars" />
    <skos:broader rdf:resource="iau93:RedStars" />
  </skos:Concept>
  <skos:Concept rdf:about='#blueVars'>
    <skos:prefLabel xml:lang='en'>blue variables</skos:prefLabel>
    <skos:definition>Variables that are brighter in the blue than in the
red.</skos:definition>
    <skos:broader rdf:resource="iau93:VariableStars" />
    <skos:broader rdf:resource="iau93:BlueObjects" />
  </skos:Concept>
  <skos:Concept rdf:about='#whiteVars'>
    <skos:prefLabel xml:lang='en'>white variables</skos:prefLabel>
    <skos:definition>Variables that are neither particularly red nor
blue.</skos:definition>
    <skos:broader rdf:resource="iau93:VariableStars" />
  </skos:Concept>
  <skos:Concept rdf:about='#blueMag'>
    <skos:prefLabel xml:lang='en'>blue brightness</skos:prefLabel>
    <skos:definition>Blue visual magnitude estimated using a comparison
LED.</skos:definition>
    <skos:broader rdf:resource="iau93:VisualMagnitude" />
  </skos:Concept>
  <skos:Concept rdf:about='#redMag'>
    <skos:prefLabel xml:lang='en'>red brightness</skos:prefLabel>
    <skos:definition>Red visual magnitude estimated using a comparison
LED.</skos:definition>
    <skos:broader rdf:resource="iau93:VisualMagnitude" />
  </skos:Concept>
</rdf:RDF>
```

The `<skos:ConceptScheme>` serves to document the vocabulary and its creator. Normally, one would place the top taxonomic entries in “top concept” elements to aid navigation, but this vocabulary is so small that this is totally unnecessary. Like NASA, one could have chosen a more obscure set of labels -- e.g. “123abc” for “redVars” -- since the formal names set by the “rdf:about” are URIs, and so really for computer consumption only. However, it's just as easy to be more explicit and human-readable. For a full set of “good practice” recommendations, see the IVOA white paper on vocabularies (<http://www.ivoa.net/Documents/REC/Semantics/Vocabularies-20091007.html#practices>), and you can download a validator for your vocabulary at [10].

The corresponding translation entries have been included here as “broader” entries, assuming that the burden on the publisher and consumer is minimal for such a small vocabulary.

This vocabulary document should be available at a clearly reachable URI, e.g.

```
http://www.blueandred.org/rdf/BlueAndRed/v9.3/BlueAndRed.skos
```

and should preferably be mirrored at the IVOA for long-term documentation purposes and ease of access. The namespace for the vocabulary terms was also given in the `xml:base` attribute in the vocabulary above. An ideal documentation strategy is to make the nominal URI of the entire vocabulary, e.g.

```
http://www.blueandred.org/rdf/BlueAndRed/v9.3
```

a normal HTTP-accessible webpage: the URI <http://www.ivoa.net/rdf/Vocabularies/IAUT93> (i.e. without the “#” separator between root address and vocabulary label) is actually a browseable list of the IAU 1993 thesaurus entries.

Step 2: Add/Parse the vocabulary to/from the VOEvent document

In order to make the “Blue and Red Variable Stars” and VOEvent vocabularies accessible, you must define your namespace and add it to the VOEvent document (or note the namespace so you can reference the items when reading a document referring to the “BlueAndRed” vocabulary):

```
<?xml version="1.0" encoding="UTF-8"?>
<voe:VOEvent
  ivorn="ivo://BlueAndRed.org/events#123457"
  role="observation"
  version="1.1"
  xmlns:voe="http://www.ivoa.net/xml/VOEvent/v1.1"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:br="http://www.BlueAndRed.org/rdf/BlueAndRed/v9.3#"
  xmlns:voe="http://www.ivoa.net/rdf/Vocabularies/VOEvent/v1.0"
  xmlns:iau93="http://www.ivoa.net/rdf/Vocabularies/IAUT93#"
  xsi:schemaLocation="http://www.ivoa.net/xml/VOEvent/v1.1
  http://www.ivoa.net/xml/VOEvent/VOEvent-v1.1.xsd">
```

This entry permits semantic references in the `<What>` section like:

```
<what>
  <Param name="br:blueMag" value="12" unit="mag" ucd="phot.mag;em.opt.B"/>
  <Param name="br:redMag" value="19" unit="mag" ucd="phot.mag;em.opt.R"/>
  <Reference uri="http://www.BlueAndRed.org/events/123456"/>
  <Description>visual magnitudes</Description>
</what>
```

in which the blue- and red-magnitudes of the observations are documented using a specialized vocabulary. Similarly one can then have entries in the `<Why>` section like:

```
<why importance="0.9" expires="2010-12-31T23:59:59">
  <Concept>br:objTyp2</Concept>
  <Description>
    More blue light than red light and not seen last year, so must be a blue variable
    star.
    Awfully big brightness difference, so maybe a brand new proto-planetary nebula
    with no Halpha?
```

```
</Description>
<Inference relation="voe:identified" probability="0.1">
  <Concept>iau93:PlanetaryNebulae</Concept>
</Inference>
</why>
```

which uses both a specialized object classification and a standardized VOEvent inference vocabulary.

Step 3: Consuming external vocabularies

The previous two steps were enough for a publisher: defining and publishing vocabularies and then using them in VOEvent documents. The trickier part is left for the consumer: how do I identify whether the new semantic information is interesting?

The first step is simply to parse the VOEvent documentation. Since the document has been parsed as XML anyway, identifying the elements with the foreign vocabulary information is straight-forward. In the example above, it may or may not be necessary to parse the <What> information: the UCD entries clearly identify the magnitudes as being bluish and reddish magnitudes. In the <Why> section, the clearest need is to be able to parse the concept information: am I interested in new blue objects and/or things potentially related to planetary nebulae? The full parsing of the <Why> information in the example above would consist of two steps: determining whether the identification is interesting and whether the inference is of relevance.

To determine if a <Why> concept is relevant and interesting, one first has to identify what is meant by “br:blueVars”. The vocabulary names are often of little direct use: while “iau93:PlanetaryNebulae” is easy to interpret, NASA’s “loc:83” is fairly obscure. Accessing the SKOS file for the “BlueAndRed” vocabulary is easily done via the namespace declaration, however, yielding a clear reference to “iau93:VariableStars” and “iau93:BlueObjects”. If your VOEvent filter is interested in blue variable objects, then it may certainly be appropriate to continue parsing this VOEvent document. If you are also interested in planetary nebulae, then the reference to the concept in the <Inference> section is very simple: even if one doesn’t keep the entire IAU 1993 thesaurus in memory, one can easily keep a sub-set of interest and test concepts and inferences against it.

Although we have talked, above, of parsing random external vocabularies, this is not something one would want to do very often. It is better regarded as a ‘compile-time’ operation in most cases.

Conclusions

Producing and consuming VOEvent documents without expressing or appreciating their full semantic context is a shame: it is really easy to place events in a clearer astrophysical semantic context, thereby helping consumers to judge the potential scientific interest and impact. VOEvent vocabularies are easily expressed, easily stored, easily found, and easily parsed. Shame on you if you don’t use them!

References

- [1] Wiktionary: <http://en.wiktionary.org/wiki/semantics>
- [2] ISO 5964:1985, Documentation — guidelines for the establishment and development of multilingual thesauri. International Standard, 1985
- [3] BS 8723-1:2005, Structured vocabularies for information retrieval
- [4] T. R. Gruber. A translation approach to portable ontology specification. Knowledge Acquisition, 5(2):199–220, 1993. do:10.1006/knac.1993.1008
- [5] NASA taxonomy: <http://nasataxonomy.jpl.nasa.gov/fordevelopers/#skos>
- [6] Hurt, Christensen, and Gauthier 2008, “Virtual astronomy metadata project” http://www.virtualastronomy.org/AVM_Version1.1May312008.pdf

- [7] Ontology of Astronomical Object Types
<http://www.ivoa.net/Documents/latest/AstrObjectOntology.html>
- [8] <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>
- [9] NASA “locations” taxonomy
<http://nasataxonomy.jpl.nasa.gov/xml/locations.skos>
- [10] Vocabulary validation
<http://code.google.com/p/volute/downloads/list>