Astronomical Data Analysis Software and Systems XVII **O9.4**
ASP Conference Series, Vol. 394, © 2008
R. W. Argyle, P. S. Bunclark, and J. R. Lewis, eds.

VIM: A Tool to Explore Your Sources

R. D. Williams

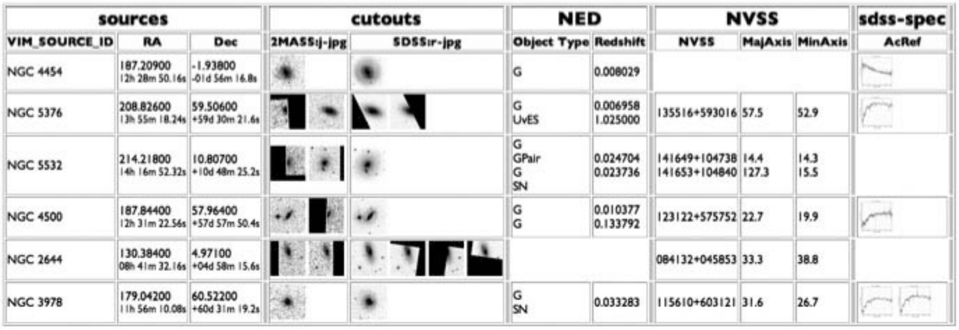
California Institute of Technology, Pasadena, CA 91125, USA

Abstract. VIM (Virtual Observatory Integration and Mining) is a web-based data retrieval and exploration application that assumes an astronomer has a list of 'sources' (positions in the sky), and wants to explore archival catalogs, images, and spectra of the sources, in order to identify, select, and mine the list. VIM does this through web forms, building a custom 'data matrix', whose rows are the uploaded source positions, and the columns show archival data – in fact any VO-registered catalog service can be used by VIM, as well as co-registered image cutouts from VO-image services, and spectra from VO-spectrum services. The user could, for example, show together: proper motions from GSC2, name and spectral type from NED, magnitudes and colors from 2MASS, and cutouts and spectra from SDSS. VIM can compute columns across surveys and sort on these (eg. 2MASS J magnitude minus SDSS g). For larger sets of sources, VIM utilizes the asynchronous Nesssi services from NVO, that can run thousands of cone and image services overnight.

One of the major thrusts of the Virtual Observatory (VO) is the resolution of sets of sources, i.e. finding digital information about a set of sources using web forms and scripts. Traditional astronomy studies one or a few objects in great depth, but the emerging science of *astroinformatics* works with populations, finding clusters, outliers, and correlations using statistics and data mining. We assume that the intention of the customer is to select 'good' sources by mining subsets of a population; to make statistical studies of a population; or to present data on these sources to humans using images and web; or to save the sources and the associated data objects; or to create a new table or catalog by federating data on these sources. VIM is intended to make a coherent picture from many sources that have been matched to many catalogs.

A principal motivation for VIM came from working with a synoptic sky survey (Palomar-Quest[1]), where a real-time pipeline creates a few to a hundred candidate transient events every night of operation, and we want to assess all of these together rapidly, in order to select the 'best' ones and encourage followup observations of those. There is a list of positions in the sky, each with an identifier that connects back to the Palomar-Quest database, and each should be compared with major catalog holdings from SDSS, 2MASS, catalogs of variables, etc, and the results presented either by row (all data for given source), or buy column (all data from given catalog). In other words, ViIM coherently summarizes the relevant data and allows selection of 'best' sources. The following definitions will provide the context for the rest of this paper.

[1] `http://www.astro.caltech.edu/g̃eorge/pq/`



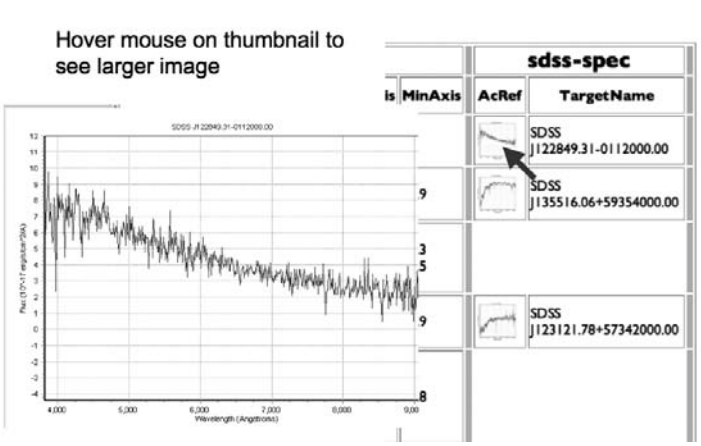
sources			cutouts		NED		NVSS			sdss-spec
VIM_SOURCE_ID	RA	Dec	2MASSj-jpg	SDSSir-jpg	Object Type	Redshift	NVSS	MajAxis	MinAxis	AcRef
NGC 4454	187.20900 12h 28m 50.16s	-1.93800 -01d 56m 16.8s			G	0.008029				
NGC 5376	208.82600 13h 55m 18.24s	59.50600 +59d 30m 21.6s			G UvES	0.006958 1.025000	135516+593016	57.5	52.9	
NGC 5532	214.21800 14h 16m 52.32s	10.80700 +10d 48m 25.2s			G GPair G SN	0.024704 0.023736	141649+104738 141653+104840	14.4 127.3	14.3 15.5	
NGC 4500	187.84400 12h 31m 22.56s	57.96400 +57d 57m 50.4s			G G	0.010377 0.133792	123122+575752	22.7	19.9	
NGC 2644	130.38400 08h 41m 32.16s	4.97100 +04d 58m 15.6s					084132+045853	33.3	38.8	
NGC 3978	179.04200 11h 56m 10.08s	60.52200 +60d 31m 19.2s			G SN	0.033283	115610+603121	31.6	26.7	

Figure 1. A screenshot from the VIM system. Sources were uploaded (left), then compared against NED, the NVSS survey, and the SDSS spectrum catalog. Cutouts were made from SDSS and 2MASS. All these data sources can be directly compared in this view of a data matrix.

- **Source**: A source is defined as a named position in the sky. Currently, VIM only accepts positions written as RA and Dec in ICRS. When sources are combined in a set, the names must be unique.
- **Catalog**: A catalog is defined as a set of points in the sky, each with a data record attached. Each data record instantiates from the same *catalog schema*.
- **Catalog service**: Associated with a catalog may be a service that matches sources against the catalog, returning data records near on the sky to each source. In many cases, this service can be implemented with a cone service, or with a VO Skynode. In short, a catalog service is a generalized cone search.
- **Data matrix**: When a set of sources is run through a set of catalog services, the result is a data matrix. For each source and catalog service, there are zero or more *catalog matches*, meaning the set of catalog entries where that source is near that catalog entry.

To start, we assume that the customer has a set of N sources, each with identifier and sky position. The customer may create these identifiers, or VIM can create them automatically (eg. a source at 21h24m32.3s +22d55m22s could be given the name “J21243230+2255220”). This identifier is then attached to any service outputs from this source. Sources can be uploaded to VIM as comma-separated values, as a VOTable with RA, Dec, ID columns, or as a URL that points to such a table, or another application can POST a VOTable source table directly to VIM.

The sources are uploaded to the server. VIM provides persistent storage on the server side; every time a new source table is uploaded, a new directory is created, called a *workbench*, that is accessible as a web page, for example: `http://myplace.edu/cgi-bin/vim.cgi?benchID=83996009453609531311235737334395`. The URL that points to a workbench can be shared with colleagues, and is, generally speaking, quite safe from public view because of the random digits in the URL that make it impossible to find without knowing. Furthermore, a write-password may be added to the workbench when it is first created, and that password is needed again to make changes to the data in the workbench. Thus a URL of a workbench can be shared, without sharing the ability to modify it. The customer can then select catalog services, either from a drop-down list of primary catalogs, or by searching in the VO Registry. A cone radius is chosen, and then


Hover mouse on thumbnail to see larger image

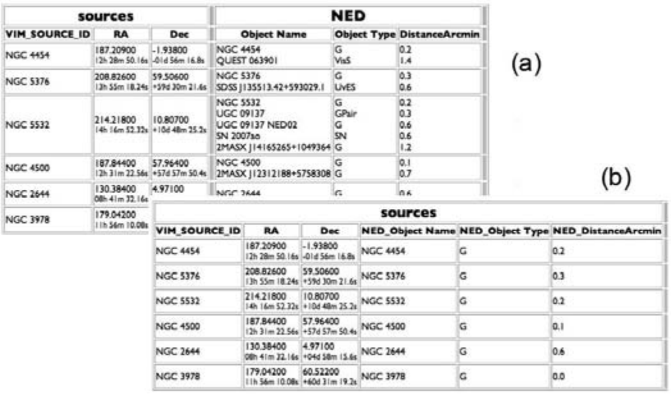
		sdss-spec	
MinAxis		AcRef	TargetName
			SDSS J122849.31-0112000.00
9			SDSS J135516.06+59354000.00
3 5			
9			SDSS J123121.78+57342000.00
8			

Figure 2. Cutouts and images of spectra are available as thumbnail images in the data display, and larger images when the mouse hovers over, and are linked to the FITS files from which the JPEG images derive.

all the sources are resolved by the service. In the current VIM, this is done by multi-threaded cone-search requests; in the future, we expect to make use of more sophisticated catalog query protocols, such as Skynode or VO Table Access Protocol.

VIM can deliver graphical information about sources in addition to numerical. Cutouts around sources are available, as well as spectra. A cutout appears in the data matrix as a URL string that points to the image; once VIM is informed that there is an image behind the URL, it will be displayed as an image. This signal is through the UCD[2] (unified content descriptor) of the table column: anything whose UCD is `obs.image.*` is assumed to be a link to an image. Images are displayed in both thumbnail and full form, where hovering the mouse over the thumbnail brings up the full image, as shown in Figure 2. VIM provides some tools for mining the data matrix. The primary tool is **Join**, where a table of matches is combined with the original source table. This can be done by selecting just the nearest matching catalog entry (proximity join), or by the “outer join”, where each source-match pair becomes a record in the source table. This tool is a kind of spatial crossmatch, and it may be argued that this is too simple; that a real crossmatch must be able to involve more complex matching based on non-spatial attributes such as color. However, as noted below, VIM provides selection methods both before and after join that allow this kind of complexity. Columns can be sorted, say on color or brightness, or anything else. As noted above, VIM distinguishes the data matrix itself (on the server) from the view of the data matrix that is given to the customer. Sorting means that only the “interesting” sources appear in that view. New columns can be computed from others by entering an arithmetic expression. For example, we may have joined 2MASS matches and SDSS matches to the source table, so that columns of magnitudes are available called *twomass_J* and *sdss_g* for the infrared J and optical g magnitudes. The expression *twomass_J* - *sdss_g*

[2] `http://vizier.u-strasbg.fr/UCD/`



sources			NED		
VIM_SOURCE_ID	RA	Dec	Object Name	Object Type	DistanceArcmin
NGC 4454	187.20900 12h 28m 50.16s	-1.93800 -01d 56m 16.8s	NGC 4454 QUEST 063901	G VisS	0.2 1.4
NGC 5376	208.82600 13h 55m 18.24s	59.50600 +59d 30m 21.6s	NGC 5376 SDSS J135513.42+593029.1	G UvES	0.3 0.6
NGC 5532	214.21800 14h 16m 52.32s	10.80700 +10d 48m 25.2s	NGC 5532 UGC 09137 UGC 09137 NED02 SN 2007ao 2MASX J14165265+1049364	G GPair G SN G	0.2 0.3 0.6 0.6 1.2
NGC 4500	187.84400 12h 31m 22.56s	57.96400 +57d 57m 50.4s	NGC 4500 2MASX J12312188+5758308	G G	0.1 0.7
NGC 2644	130.38400 08h 41m 32.16s	4.97100	NGC 2644	G	0.6
NGC 3978	179.04200 11h 56m 10.08s				

sources					
VIM_SOURCE_ID	RA	Dec	NED_Object Name	NED_Object Type	NED_DistanceArcmin
NGC 4454	187.20900 12h 28m 50.16s	-1.93800 -01d 56m 16.8s	NGC 4454	G	0.2
NGC 5376	208.82600 13h 55m 18.24s	59.50600 +59d 30m 21.6s	NGC 5376	G	0.3
NGC 5532	214.21800 14h 16m 52.32s	10.80700 +10d 48m 25.2s	NGC 5532	G	0.2
NGC 4500	187.84400 12h 31m 22.56s	57.96400 +57d 57m 50.4s	NGC 4500	G	0.1
NGC 2644	130.38400 08h 41m 32.16s	4.97100 +04d 58m 15.6s	NGC 2644	G	0.6
NGC 3978	179.04200 11h 56m 10.08s	60.52200 +60d 31m 19.2s	NGC 3978	G	0.0

Figure 3. Joining a match table to the source table. In (a) the sources on the left are matched to NED on the right, with one source having five matches. In (b) the nearest match has been chosen, extending the source table with NED data. Other Join methods are available in addition to nearest.

could be used to generate an infrared-optical color for each object. The Select tool uses an arithmetic predicate to remove sources or matches that satisfy the criterion. For example, we can remove faint objects by running Select with the criterion $sdss\_g > 22$.

In order to scale up to very large source collections, VIM can utilize the asynchronous Nesssi[3] services from NVO, in particular the service for creating cutouts. The source list is uploaded and a set of VO image servers selected, then all those cutouts made, each reprojected to a common pixel plane, so that all the cutouts 'line up', and both FITS and JPEG representations kept and linked to each other. The resulting table of links can be added to the VIM data matrix. For efficiency, VIM also provides a caching tool so that all the links can be resolved and the cutouts copied to the VIM workbench.

What VIM does is to take a set of *sources* and a set of *catalogs* and look up each source in each catalog in order to create a data matrix. VIM is part of the emerging NVO portal architecture, which presents a collection of applications, each with the same paradigm of sources, catalog services, and a data matrix. Other components of the Portal allow sources to be derived from a catalog, or sources to be uploads in different formats; for catalog services to be found from the VO registry, or for catalogs to be found because they overlap well with a source list.

Acknowledgements. We are grateful to the National Science Foundation for support of this work through the National Virtual Observatory project, and to the UK Astrogrid project (M. Taylor) for the Stilts package.

[3] http://us-vo.org/nesssi