Baltic Astronomy, vol. 21, 301–308, 2012

SCIENCE INITIATIVES OF THE US VIRTUAL ASTRONOMICAL OBSERVATORY

Robert J. Hanisch[1,2,3]

[1] *Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218 USA; hanisch@stsci.edu*

[2] *Virtual Astronomical Observatory LLC, 1400 16th Street NW, Suite 730, Washington, DC 20036 USA; director@usvao.org*

[3] *On behalf of the VAO project team*

Received: 2012 February 13; accepted: 2012 May 18

Abstract. The United States Virtual Astronomical Observatory program is the operational facility successor to the National Virtual Observatory development project. The primary goal of the US VAO is to build on the standards, protocols, and associated infrastructure developed by NVO and the International Virtual Observatory Alliance partners and to bring to fruition a suite of applications and web-based tools that greatly enhance the research productivity of professional astronomers. To this end, and guided by the advice of our Science Council (advisory committee), we are focusing on five science initiatives in the first two years of VAO operations: (1) scalable cross-comparisons between astronomical source catalogs, (2) dynamic spectral energy distribution construction, visualization, and model fitting, (3) integration and periodogram analysis of time series data from the Harvard Time Series Center and NASA Star and Exoplanet Database, (4) integration of VO data discovery and access tools into the IR AF data analysis environment, and (5) a web-based portal to VO data discovery, access, and display tools. We are also developing tools for data linking and semantic discovery, and have a plan for providing data mining and advanced statistical analysis resources for VAO users. Initial versions of these applications and web-based services are being released over the course of the summer and fall of 2011, with further updates and enhancements planned for throughout 2012 and beyond.

Key words: virtual observatory, cross-matching, spectral energy distributions, periodograms, data-mining

1. INTRODUCTION

The US Virtual Astronomical Observatory (VAO)[1] is the operational successor to the National Virtual Observatory (NVO) development project. While a modest level of work continues on data access protocols and related standards (in collaboration with the International Virtual Observatory Alliance, IVOA),[2] the primary focus of the VAO is on providing science capabilities in a robust operational en-

[1] http://www.usvao.org

[2] http://www.ivoa.net

vironment. The science capabilities selected for implementation are determined in consultation with the VAO Science Council through an annual review process (Fabbiano et al. 2011).

2. DATA DISCOVERY PORTAL

The VAO Data Discovery Portal is the primary Web-based application for exploration and access to the thousands of Virtual Observatory (VO)-accessible data collections and catalogs worldwide. The first release of the VAO Data Discovery Tool, in July 2011, provides a simple, "one box" query interface in which an object name or position may be entered. The interface also provides filters, or "facet" to allow for quick subsetting of search results. All functions are web browser-based, so that no software installation is required. A number of VO applications incorporate data discovery capabilities (e.g., Aladin, DS9, TOPCAT, VOSpec, etc.) but in the VAO we are striving to integrate several approaches to data discovery – search by name, by position, by object type or class, by bibliographic references, by provenance – into a common environment.

Results tables can be sorted alphabetically or numerically simply by clicking on the column heading, and columns can be reordered by just dragging the column to a different location. Pull-down menus appear next to column headings that allow users to show additional metadata columns, or hide those that do not interest them. The user can examine the records in any particular resource, download them for local analysis, and drill down to the actual data set for download or automatic display.

As we continue work in the VAO, the Discovery Portal will remain one of the key science initiatives. We will enhance the interface, including searching by scientific topic and the ability to quickly find images covering a wide area of the sky, as well as improve the overall usability of the interface. In addition, we plan to focus on two key themes: *interconnectability* and *overview and drilldown*.

With overview and drilldown, we will provide a means for the user to gain a quick understanding of what kinds of data are available as well as how to focus in on the data of interest quickly. The user should be able to not only know what data collections are available but also choose which collections to focus on. This choice may be simple, as in searching just the major collections across the wavebands versus searching everything known to the VO, or it could be more customized, by searching a selected list of collections. A comprehensive approach means that searching cannot be simply position-based; it needs to also support searching by topic, such as object type and astronomical phenomenon. Building on the progress from our first-year study on literature-based discovery, we want to integrate semantic searching techniques into the overall discovery process. When drilling-down into specific collections and datasets, the user must be able to inspect and understand the detailed metadata and provenance information in order to select collections and databases of interest.

Interconnectability is about having the Discovery Portal work seamlessly with other tools that the user employs, particularly on the desktop. With data selected, the user will be able to pull those data into other tools and environments for further manipulation and analysis. For example, upon locating an interesting catalog, the user should be able to switch readily to the TAP client for advanced catalog searching. This theme is important also to our overall desktop integration strategy.

As we focus on these two themes, the following capabilities will be given highest priority:

• Provide information about what data collections and services are known to the Discovery Portal prior to actually issuing a search, so the user knows in advance where data will be sought. Allow users to discover specific types of data collections and to restrict data searches to those collections.

• Integrate literature-based searching into the Discovery Portal, so searches can begin with topics, papers, and/or authors in addition to object names and positions.

• Expose data provenance and metadata more thoroughly, using column labels associated with IVOA data models and data access protocols, rather than only with the column labels intrinsic to the service or collection. This enables easier understanding of the commonalities among diverse data collections and services.

• Improve the embedded help features (context-sensitive help, help messages embedded in the tool rather than in an external Web page) and integrate or provide access to the Table Access Protocol (TAP) Client from the Discovery Portal.

• Provide a "shopping cart" capability so the user can mark results for temporary storage and later downloading, and provide various output format options (VOTable, FITS table, CSV, plain ASCII). This will also allow the user to store results in a VOSpace (a shareable virtual storage area) rather than requiring a download to the users local storage.

• Incorporate the Simple Applications Messaging Protocol (SAMP) so data discovery results can be more easily shared with other VO-enabled applications (e.g., Topcat, Aladin, DS9). This avoids the need to download a results table and read that table into the other application manually.

3. SCALABLE CROSS-MATCHING FACILITIES

The analysis and interpretation of multi-wavelength data, especially from surveys, relies on the comparisons of source properties (flux, morphology) from different bandpasses, instruments and telescopes. The first step in such analysis is to perform a spatial cross-match between source lists or catalog entries based on position and a region of interest, such as a position uncertainty ellipse, or a PSF size. This approach is far from straightforward, however. Unless the source catalogs and object lists are based on data that are of very similar spatial resolution, similar sensitivity, and fully overlapping spatial coverage, it is difficult to assess whether an apparent match is of physical significance (originating from the same astrophysical entity) and whether a lack of a match is also significant (rather than a sensitivity mismatch or a region lacking coverage in one or another catalog). Furthermore, astrophysical objects manifest themselves in profoundly different ways in different parts of the spectrum. A radio galaxy may show bright, extended lobes of emission with faint/unremarkable nuclear emission, but be physically associated with a distant elliptical galaxy. Matching the radio lobes to optical objects would give spurious results. Also, as deep optical/infrared surveys have shown, the light of galaxies at increasing redshifts eventually moves out of the optical bandpasses. This is a significant effect for inferring the distances of the galaxies and indicates that negative cross-matches are as important as positive ones.

We have developed a simple spatial cross-comparison tool capable of matching an object list of 1 M positions against a survey catalog of 1 B positions within 1–2

minutes. This tool computes "matches" simply on the basis of a user-specified positional coincidence, and can therefore produce multiple matches for a given position. The tool makes no assertion as to the physical validity of one match over another. However, it operates quickly, allows users to upload their own object lists, chooses target catalogs from among all catalogs registered with the VO that have positional information, and operates synchronously for small target lists and asynchronously for large target lists. We also developed a prototype Bayesian cross-matching facility that utilizes positional uncertainty estimates and photometric error estimates to determine the likelihood that a cross-match is of physical significance (Budavari & Szalay 2008).

Determining the validity of cross-matches, especially between widely separated spectral bandpasses or between observations with substantially different spatial resolution, is something that cannot currently be done without scientific reasoning, and criteria may well depend on the type of question being asked of the data. Therefore, we will focus efforts on a few enhancements to the tool that provides cross-match candidates based on positional coincidence.

4. BUILDING AND ANALYZING SPECTRAL ENERGY DISTRIBUTIONS

Spectral energy distributions (SEDs) are one of the most fundamental tools for understanding the physical nature of astronomical objects. SEDs are typically constructed by assembling disparate photometric and spectroscopic measurements from the literature and online catalogs and databases. The NASA/IPAC Extragalactic Database (NED) constructs SEDs for extragalactic objects through systematic scanning of the astronomical literature for new measurements, entering those measurements and their associated metadata (filter or bandpass, aperture, units of measurement) into the database, and then converting the measurements to a common set of units. It is difficult, however, to automatically correct for differences in aperture, and there are always problems arising from errors in the published metadata. Thus it is not uncommon to see SEDs with flux measurements in the same bandpass that are apparently inconsistent. Are these the result of aperture differences, metadata errors, or more interesting astrophysically time-varying phenomena?

We have completed a core toolkit for building, displaying, and fitting SEDs based on an extensive requirements document (D'Abrusco & McDowell 2010). The *SED Importer* allows a user to retrieve SED spectro-photometry measurements from NED and to augment those data with their own measurements, or measurements from other catalogs (including accepting inputs from any SAMP-enabled application such as Topcat). The resulting SED is read into *Iris*, which is an integrated display and fitting tool based on STScI's *Specview* (Busko 2002) and SAO's *Sherpa* (Doe et al. 2007).

Priorities for further development of *Iris* include:

- expanding the options for data gathering in the SED builder,
- providing support for rebinning photometry points,
- giving greater control to the user in interactively examining the provenance of individual points in the SED and deciding to include or exclude certain points,
- providing additional fitting functions including templates for stellar population analysis, and
- incorporating corrections for extinction and redshift.

5. TIME DOMAIN ASTRONOMY

The time domain is often referred to as the last frontier in astrophysics. Although much is known about time-variable phenomena, from solar and stellar pulsations to gamma-ray bursts, our sampling of the data in the time domain is terribly incomplete. A number of new time domain surveys are now in progress or being planned, most notably with the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) and the Large Synoptic Survey Telescope (LSST). These programs will undoubtedly give rise to discoveries of previously unknown time-variable phenomenon, and will enrich our understanding of already-known phenomenon by increasing our sample sizes by orders of magnitude.

The VAO connection to time d omain studies is primarily in two areas: distribution and management of transient event notices (VOEvent) and discovery, access, and analysis of time series data. Our VOEvent infrastructure is well developed and based on an IVOA standard. Dozens of observatories worldwide are participating in the VOEventNet network, with transient event notices automatically streaming onto the Internet and VOEvent subscribers listening in for events that interest them. SkyAlert.org (a project funded by NSF and led by VAO colleagues at Caltech and NOAO) provides a clearinghouse of VOEvent notices and allows individuals to subscribe to and/or monitor the events coming from various event streams and feeds. As the SkyAlert project funding comes to a close in 2012, the VAO will be working to establish a long-term support infrastructure for dissemination and preservation of event notices. Even though LSST will not come into operation for many years, other telescopes are already producing event streams of sufficient quantity and of great interest to the U.S. astronomical community that a robust system must be maintained and expanded. LSST and VAO have recently agreed on an Memorandum of Understanding (MOU) that will assure the research community of access to LSST data through VO protocols, and assure LSST that VO support will be provided in critical areas. A particular challenge concerning transient event notices is that LSST anticipates distributing as many as 100 000 notices per night.

There are several significant repositories of time series data, such as the Time Series Center (TSC) at Harvard, the NASA/IPAC/NExScI Star and Exoplanet Database (NStED), the Kepler mission archive at STScI/MAST, the Catalina Real-time Transient Survey (CRTS), and the COROT mission database from the European Space Agency. Thus far we have completed a virtual integration of TSC and NStED, which enabled NStEDs Periodogram Service to be used on TSC data sets. Kepler and CoRoT data are already available through NStED's Periodogram Service.

We will broaden the integration to include at least the Catalina Real-time Transient Survey (CRTS, with $\sim$500 M light curves) and explore potential collaborations with other research groups in the community active in the analysis of time series data.

6. DATA LINKING AND SEMANTIC DISCOVERY

There is no application in astronomy today that is able to combine a bibliographic search with a data search based on observational parameters (such as instrument, wavelength, object classification, etc.) and semantic information about astronomical objects into a user-friendly interface. Searches such as "find all pa-

pers describing UV imaging observations of clusters of galaxies" or "find all fields of view described in Jones et al. (2010) that have been observed by HST and Chandra" lead to maximum re-use of existing data, better efficiency in allocation of rare resources (new observing time), and maximum credibility of the research record itself. But for such a capability to be useful, it must be low latency in the face of huge amounts of metadata so the user can work in an exploratory mode, and when done, switch to a high bandwidth mode to be able to make downloads automatically from the VAO infrastructure.

We have developed a prototype discovery tool based on the astronomical literature, an ontology of astronomical terms, and links to the underlying data in two major data centers (Chandra, MAST). A key feature of the discovery tool is "faceted browsing" in which query results may be instantly filtered by various attributes (authors, publication dates, data type, data source, etc.). Our goal is to extend the scope of content in this tool and to more tightly integrate it with the general data discovery portal.

The VAO will also work with the American Astronomical Society, the Astrophysics Data System, NED, and major international data centers such as the Centre de Données astronomiques de Strasbourg (CDS) to develop a process for data capture and validation as part of the publication process, and for long-term storage of the digital data objects appearing in graphical form in journal articles. Our goal is to have the digital data represented in scholarly publications captured, curated, and preserved, linked from the journals, and independently discoverable and downloadable through VO interfaces.

7. DESKTOP TOOL INTEGRATION

The astronomical research community utilizes a variety of desktop computer environments and tools to do their data reduction and analysis. There are large, general purpose systems that originated from and are supported by organizations within astronomy, such as the Image Reduction and Analysis Facility (IRAF), Astronomical Image Processing System (AIPS), Common Astronomy Software Applications (CASA), Chandra Interactive Analysis of Observations (CIAO), and ESO-MIDAS. There are also commercial packages, most notably IDL, and public domain environments like Python and R, that have wide or growing use, owing to the ease of prototyping and development. (Python is also becoming an integral part of environments such as IRAF and CASA.) And there are a myriad of more special-purpose applications, such as DAOPhot and SExtractor for photometry, that are in wide use and are generally accepted as tried-and-true tools. Of immediate concern is to bring VO-based capabilities into the environments and applications that astronomers already know, and making it as easy as possible to share information among these tools.

There is already an IVOA standard for basic applications interoperability called SAMP, the Simple Applications Message Protocol. SAMP allows several independent applications running on the user's desktop to communicate with each other, so that, for example, a selection of objects in an image display application, such as Aladin, automatically updates a scatter plot in a graphics application, such as Topcat. We will incorporate SAMP into the Data Discovery Portal and other VO-based applications.

We have completed the VO/IRAF integration project. IRAF is in use by perhaps 5,000 or more astronomers worldwide. By making VO capabilities intrinsic

to IRAF (registry searches, use of remote data as if it were local, communication between IRAF tasks and other VO applications via SAMP), these astronomers can expand their VO horizons from within a familiar environment. Within the desktop environment, IRAF use SAMP messaging to exchange data and commands with other VO applications, creating an integrated suite of tools more powerful than the individual components. Calls to remote VO services are used to query for additional data or upload local data for analysis by a Web service. IRAF is able to query/access remote data using standard HTTP protocols and make use of the XML documents and FITS files returned using native interfaces. Extensions to the scripting environment allow for creation of new science tasks based on these enhanced capabilities.

Our near-term goal is to bring the VO to astronomers, not to require astronomers to come to the VO and think they need to learn an entirely new way of working. To this end we will expand the VO/IRAF concept through collaborations with the organizations that support CASA (National Radio Astronomy Observatory), CIAO (Smithsonian Astrophysical Observatory), and PyRAF (Space Telescope Science Institute), though as these efforts would rely on non-VAO sponsored work within those organizations, they are long-term initiatives. We have very basic VO interfaces for IDL (developed under NVO), and these will need to be updated and repackaged for distribution from VAO. We are investigating strategies for R/VO integration with R experts in the community.

8. DATA MINING, STATISTICAL ANALYSIS AND VISUALIZATION

Data mining, multivariate statistical analysis, and the emerging field of astroinformatics are all aimed at distilling information from large, multi-dimensional, heterogeneous data sets. This includes identifying unique or unusual classes of objects, estimating correlations, computing the statistical significance of a fit to a model in the presence of missing data or data with upper limits. The computational challenges can be enormous, but so can the barriers to using and understanding the tools of the trade. The more sophisticated statistical methods are not often used in astronomy, and determining which statistical test to use for what situation requires learning a complex jargon.

We undertook a study to determine where adaptation and integration of data mining, statistical analysis, and visualization tools will be most effective. We evaluated the most widely used applications and toolkits, both within and outside of astronomy. We believe the most cost-effective strategy for introducing these tools to the community is to work with the Data Mining and Exploration (DAME) collaboration between the University of Naples, Italy, Capodimonte Astronomical Observatory in Naples, Italy, and the California Institute of Technology whose web site and tools are already quite well developed.[3] Caltech collaborators include VAO team members, and two former graduate students at Naples who helped develop DAME are now participating in VAO activities at SAO. In order to more closely link DAME services with the VO, it will be important to enable DAME to access data through a VOSpace storage area.

[3] http://dame.dsf.unina.it

9. CONCLUSIONS

The US VAO endeavors to provide a research environment that facilitates access to diverse and distributed data, makes it easy to compare such data, and enables astronomers to interact efficiently and effectively with large and complex data collections and services. We are eager to hear from users about their experiences and to get suggestions for improvements and additions to our tools. You can access the VAO and submit questions and ideas through the VAO web site at http://www.usvao.org/.

ACKNOWLEDGMENTS. The developments mentioned here are the work of the VAO product development team, with contributions from the California Institute of Technology, the Infrared Processing and Analysis Center (Infra-Red Science Archive and NASA Extragalactic Database), the Johns Hopkins University, NASA's Goddard Space Flight Center, the National Center for Supercomputing Applications at the University of Illinois, the National Radio Astronomy Observatory, the National Optical Astronomy Observatory, the Smithsonian Astrophysical Observatory, and the Space Telescope Science Institute. The VAO is a member of the International Virtual Observatory Alliance. The US VAO program is sponsored by the National Science Foundation and the National Aeronautics and Space Administration.

REFERENCES

Budavari T., Szalay A. S. 2008, ApJ, 679, 301
Busko I. 2002, ADASS XI, ASPC, 261, 120
D'Abrusco R., McDowell J. and the VAO Team 2010, arXiv:1012.5733v1
Doe S. et al. 2007, ADASS XVI, ASPC, 376, 543
Fabbiano G. et al. 2011, arXiv:1108.4348v2