DATA STANDARDS FOR THE INTERNATIONAL VIRTUAL OBSERVATORY

R. J. Hanisch

Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218.
Email: hanisch@stsci.edu

ABSTRACT

A primary goal of the International Virtual Observatory Alliance, which brings together Virtual Observatory Projects from 16 national and international development projects, is to develop, evaluate, test, and agree upon standards for astronomical data formatting, data discovery, and data delivery. In the three years that the IVOA has been in existence, substantial progress has been made on standards for tabular data, imaging data, spectroscopic data, and large-scale databases and on managing the metadata that describe data collections and data access services. In this paper, I describe how the IVOA operates and give my views as to why such a broadly based international collaboration has been able to make such rapid progress.

Keywords: Standards, Data formats, Metadata, XML, Astronomy, Virtual observatory

1 INTRODUCTION

Astronomy is facing a data avalanche. Large digital sky surveys have become the dominant source of data in astronomy, now totaling well over 100 Terabytes in volume and growing rapidly. There are many examples, such as the Sloan Digital Sky Survey (SDSS), the Two Micron All Sky Survey (2MASS), the Palomar Digital Sky Survey (DPOSS), the Guide Star Catalog (GSC), Faint Images of the Radio Sky (FIRST), the National Radio Astronomy VLA Sky Survey (NVSS), the ROSAT All-Sky-Survey (RASS), the Infrared Astronomical Satellite (IRAS), The Quest for Excellence for Suppliers of Telecommunications (QUEST), and the Galaxy Evolution Explorer (GALEX). (A web search on any of these will lead to websites describing the missions and surveys.) Observatory archives provide vast collections of pointed observations, and future synoptic survey telescopes such as LSST will provide complete images of the observable night sky every 3-4 days. Complementing these vast data archives are fully electronic journals and astronomy digital libraries such as the Astrophysics Data System (http://adswww.harvard.edu), NASA Extragalactic Database (http://ned.ipac.caltech.edu), and CDS, the astronomy data center in Strasbourg, France, (http://cdsweb.u-strasbg.fr). Thus, astronomers are faced with data sets that are orders of magnitude larger, and more complex, than anything they have dealt with in the past. These data sets must be compared with models and theoretical simulations that are often of equal or even larger scale. The "new astronomy" emphasizes the analysis of multi-wavelength data from billions of objects, from which we expect to discover significant patterns previously hidden in smaller, biased samples. Navigation of this vast sea of data requires new approaches to data discovery, data access, and data delivery based on internationally agreed standards.

The *virtual observatory* (VO) framework, now in development by sixteen national and international projects, provides the interface standards for such navigation (Quinn et al., 2004). In addition, virtual observatory projects are providing high-level applications to work with standard VO data products and enabling astronomical researchers and data centers to provide their data to the VO with relative simplicity. Through these developments, the virtual observatory becomes a catalyst for worldwide access to astronomical observations and an essential part of the research astronomer's toolkit. It is important to emphasize, however, that the VO is not a centralized data repository or a data quality enforcement organization. The VO is the

glue that links diverse data collections together and facilitates discoveries that can only arise through data diversity. The VO complements new telescopes—on the ground and in space—by increasing the scientific return on these facilities and guiding the construction of new facilities to areas of greatest need.

2 NATIONAL AND INTERNATIONAL COOPERATION

In the United States, the National Virtual Observatory development project (http://www.us-vo.org) is funded by NSF's Information Technology Research Program. The US NVO team includes 17 organizations, including national astronomy data centers, telescope operations centers, astronomy research universities, computer science research organizations, and supercomputer centers. Now in the fourth year of a five-year development project, the US NVO has worked to establish standards for data discovery and data access and has shown how these standards can be used to develop powerful research applications. NSF and NASA are expected to jointly fund a long-term NVO Facility beginning in 2006.

The US NVO project is a co-founder of the International Virtual Observatory Alliance (http://www.ivoa.net). The sixteen member projects of the IVOA have agreed upon a standards process that is modeled after the World Wide Web Consortium (W3C) (2004). This process documents progress through three-stages, from working draft, to proposed recommendation, to recommendation. An optional fourth stage (not yet implemented) takes IVOA recommendations to the International Astronomical Union, the highest international body in astronomical research, for endorsement.

3 STANDARDS FOR THE VIRTUAL OBSERVATORY

3.1 Data Formats

Data standards in astronomy began with the Flexible Image Transport System (Wells, Greisen, & Harten, 1981) (Hanisch et al., 2001), best known as FITS, over 25 years ago. FITS has been adopted by every major data producing organization and is now both the archival format and run-time format of choice in virtually all data analysis software systems that are widely used in astronomical research. FITS has grown from an image exchange mechanism to something much more general, with support for text-based tables, binary tables, and spectral data. A series of metadata conventions provides for the encoding of complex "world coordinate systems" and the associated geometric projections needed to map data from the sphere of the night sky to the planes of maps and image displays (Greisen & Calabretta, 2002). FITS is fully endorsed by the IAU. Because FITS is so widely used, proposed modifications and extensions are evaluated extremely carefully, and changes are made slowly. This is viewed in the community as both a benefit and a curse! Despite the universal use of FITS, it remains primarily a syntactic standard. The rules for constructing valid FITS files, and for representing metadata, are clear, but with the exception of coordinate system specifications, there are few agreements about semantics. Data exchange and, most importantly, interoperability are limited by the lack of semantic standards.

The virtual observatory projects sought to address the shortcomings of FITS, and exploit the strengths of the industry-standard Extensible Markup Language (XML) (http://www.w3.org/XML/), by developing a new format called VOTable (Ochsenbein et al., 2004). VOTable provides for the exchange of tabular information (both text and binary) and for standard semantic mark-up of table contents, i.e., column headings, through use of Unified Content Descriptors (UCDs) (Derriere et al., 2004). In VOTables, columns can have whatever label data providers choose, but columns are also assigned a UCD from the UCD dictionary. The UCD structure and dictionary contents are also agreed within the IVOA. VOTables are defined structurally with an XML schema and can be easily transformed to other formats with XML style sheets.

VOEvent is a recently developed standard for representing the time and location of transient events in astronomy (Seaman et al., 2005). Rapid dissemination of the position on the sky, approximate brightness, and other information (either observed or inferred, such as "we think this is a supernova") is essential for some areas of astrophysical research. VOEvent notifications can be acted upon by robotically controlled

observatories in order to monitor rapid changes and plan follow-up observations. VOEvent is a specialized VOTable, defined by its own schema.

3.2 Data Discovery

The Google mode of information discovery has serious limitations for astronomical data, as most images have little or no text-based descriptions aside from object names (and astronomers have, in some cases, dozens of different names for the same objects). Thus, a major thrust of the virtual observatory has been in the area of defining resource metadata (Hanisch, 2004). Resource metadata describes the organizations and the data collections they provide, along with the associated data delivery services and computational services. For the VO we began with the Dublin Core (2004) metadata definitions and to these added astronomy-specific additions. Our metadata is encoded in XML using an IVOA-agreed schema, and we have further agreed upon mechanisms for accommodating subdiscipline-specific extensions. Resource metadata are stored in *registries*, and registries are kept synchronized through the exchange of metadata records using the Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH) (Open Archives Initiative, 2002). The registries are searchable through web interfaces and software-accessible through web services Simple Object Access Protocol (SOAP) (http://www.w3.org/TR/SOAP/) and Web Service Definition Language (WSDL) (http://www.w3.org/2002/ws/). Two important components of the resource metadata are *space-time coordinates* and *identifiers*. Space-time coordinates (STC) (Rots, 2005) provide a complete XML description of the location of an astronomical object (position on the celestial sphere, distance if known, time-frame in which an observation was made, location of the observatory, etc.), and identifiers provide a unique and long-lived label for a resource or service. VO identifiers are essentially universal resource identifiers (URIs) with an agreed upon substructure (Plante et al., 2005).

3.3 Data Access

Once data collections of interest to the researcher are discovered, the next matter is to locate and retrieve information (catalog entries, images, spectra) from these collections through standard interfaces, rather than having to craft custom queries and requests for every collection (the situation in the past). The simplest possible interface is a *cone search* so-called because the query is based on a sky position (right ascension, declination) and a radius about that position, i.e., a cone whose vertex is the observer and whose base is a circular region on the sky. A cone search request to a data collection yields a response that is encoded as a VOTable. The rows of the table list objects or observations that fall within the circular region.

The first generalization of the cone search is the *Simple Image Access Protocol* (SIAP) (Tody & Plante, 2004). In the SIAP, the query syntax is extended to permit a specification of the size of the region desired. The response is again a VOTable whose rows describe matching images. One column of the table provides a link to the image data itself, either in its FITS format representation or in one of the various industry-standard graphics formats.

The second generation of the cone search is the *Simple Spectral Access Protocol* (SSAP) (Dolensky & Tody, 2004). The SSAP query interface provides for additional query parameters (most importantly, to allow specification of the wavelength region of interest), and the VOTable response can either point to external spectral data files or encode the spectra in-line in simple (wavelength, flux) columns.

The cone search protocol is too simplistic for use with the large survey catalogs now being produced in astronomy. These catalogs are stored in relational databases such as SQLServer (http://www.microsoft.com/sql/), Oracle (http://www.oracle.com), or Sybase (http://www.sybase.com), and it is not unusual for each entry in the catalog to have tens or even hundreds of attributes. Access to such resources requires agreement on a VO query language, called Astronomy Data Query Language (ADQL) (Ohishi & Szalay, 2005), and a standard database interface that understands ADQL, converts queries to the internal database system, and formats responses in a standard way to allow comparison and cross-matching. The VO solution here is OpenSkyNode, the standard interface wrapper for relational databases (O'Mullane & Ohishi, 2005). The OpenSkyQuery portal allows users to generate ADQL queries to any registered OpenSkyNodes and to find positional cross-matches between distributed databases. Astronomy Data

Query Language (ADQ) is Structured Query Language (SQL) with astronomy-specific extensions (e.g., to support the notion that celestial coordinates such as right ascension and declination always need to be considered together).

3.4 Web Services

Thus far, the VO data access protocols have been implemented as basic HTTP GET services. However, to enable the development of sophisticated software that utilizes many distributed collections and services, we are moving toward building web services interfaces (SOAP, WSDL) throughout the VO. This requires defining a basic service profile (registration, aliveness tests, usage logging, execution IDs), supporting asynchronous services, and supporting authentication and authorization. Development is drawing from industry and grid standards in these areas.

4 THE STANDARDS PROCESS

As mentioned briefly in paragraph 2, the IVOA standards process is modeled on that of the World Wide Web Consortium, though somewhat simplified owing to the smaller size of the IVOA collaboration and well-established technical collaborations. The IVOA charters Working Groups in areas where standards are needed. At this time these include Resource Registry (resource metadata, identifiers), Content Description (UCDs), VOTable, VOEvent, VO Query Language (ADQL), Data Access (SIAP, SSAP), Grid and Web Services, and Data Models (STC). The Working Groups work through e-mail distribution lists, a TWiki collaborative web site, and semi-annual technical meetings. Leadership of the Working Groups is shared amongst the international VO projects so that no projects become dominant.

The standards process itself recognizes three levels of documents: Working Drafts, Proposed Recommendations, and Recommendations (Hanisch & Linde, 2003). Promotion from suggestion to recommendation follows these steps:

1. Working Group prepares Working Draft (version ≥1.0) and submits to Document Coordinator for posting in the IVOA document collection.
2. Working Group reviews the Working Draft. Two reference implementations of any associated software are recommended.
3. The Chair of the Working Group, with consent of the WG, promotes the document to a Proposed Recommendation and submits it to the Document Coordinator for posting in the IVOA document collection.
4. The Chair of the Working Group issues a formal Request for Comments (RFC) to the e-mail distribution list interop@ivoa.net. The RFC and all comments must be logged on a TWiki page (http://www.TWiki.org/cgi-bin/view/) whose URL is given in the RFC. A minimum comment period of 4 weeks must be allowed.
5. The Working Group Chair responds to comments on the TWiki page. If comments lead to significant changes to the document, the status reverts to Working Draft (back to Step 1).
6. If comments are addressed to the satisfaction of the WG Chair and WG members, the WG Chair requests the IVOA Chair to submit it to the Executive Committee for approval.
7. The Executive Committee is polled by the IVOA Chair to ascertain if there is consensus for promotion to Recommendation.
8. If yes, the IVOA Chair reports on approval to the WG Chair and asks the Document Coordinator to update the document status to Recommendation. If no, the concerns of the IVOA Executive need to be resolved and a new poll taken, or if serious revisions are required, the document reverts to Step 1.
9. The IVOA Executive may propose to the IAU Commission 5 that IVOA Recommendations be endorsed as IAU Standards.

5 CONCLUSION

To date, the IVOA standards processes have worked very well. This success stems from a strong, bottom-up motivation to establish a single, global set of standards for the VO. Since there is no exchange of funds among the international VO partners, technical and scientific motivation is the undisputed driver. Technical and political leadership of the IVOA and its Working Groups rotates in order to make sure no projects become overly influential and that all collaborators have a voice. The IVOA initiative is also "right-sized" - large enough that the effort can be distributed, but small enough that communications are reasonably efficient. Finally, the IVOA borrows liberally from the standards developed elsewhere in the IT community, adopting and adapting rather than reinventing.

6 ACKNOWLEDGEMENTS

The work reported here is based on the efforts of the US National Virtual Observatory project and our many international collaborators in the International Virtual Observatory Alliance. The US NVO project is supported by the National Science Foundation under Cooperative Agreement AST0122449 to The Johns Hopkins University.

7 REFERENCES

Derriere, S., et al. (2004) UCD in the IVOA Context. *ADASS XIII, ASP Conference Proceedings, 314*, 315-318. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/UCDlist.html

Dolensky, M., & Tody, D. (2004) Simple Spectrum Access Specification, Version 0.1. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/internal/IVOA/IvoaDAL/ssa-v01-MD.doc

Dublin Core Metatdata Initiative (2004) Dublin Core Metadata Element Set, Version 1.1. Retrieved from the World Wide Web, September 16, 2006: http://dublincore.org/documents/dces/

Greisen, E. W., & Calabretta, M. R. (2002) Representations of world coordinates in FITS. *A&A 395*, 1061-1075.

Hanisch, R. (2004) Resource Metadata for the Virtual Observatory, Version 1.01. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/REC/ResMetadata/RM-20040426.html

Hanisch, R. J., et al. (2001) Definition of the Flexible Image Transport System (FITS). *A&A 376*, 359-380.

Hanisch, R. J., & Linde, A. E. (2003) IVOA Document Standards, Version 1.0. Retrieved from the World Wide Web, September 16, 2006: http://ivoa.net/Documents/latest/DocStd.html

Ochsenbein, F., et al. (2004). VOTable: Tabular Data for the Virtual Observatory. *Toward an International Virtual Observatory* 118-123. Garching, Germany. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/ latest/VOT.html

Ohishi, M., & Szalay, A. (2005) IVOA Astronomical Data Query Language, Version 1.01. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/ADQL.html

O'Mullane, W., & Ohishi, M. (2005) IVOA SkyNode Interface, Version 1.01. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/SNI.html

Open Archives Initiative (2002) The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from the World Wide Web, September 16, 2006: http://www.openarchives.org/OAI/openarchivesprotocol.html

Plante, R., et al. (2005) IVOA Identifiers, Version 1.10. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/IDs.html

Quinn, P., et al. (2004) The International Virtual Observatory Alliance: recent technical developments and the road ahead. *SPIE 5493*, 137-145.

Rots, A. (2005) Space-Time Coordinate Metadata for the Virtual Observatory, Version 1.21, Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/STC.html

Seaman, R, et al. (2005) Sky Event Reporting Metadata (VOEvent). Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/Documents/latest/VOEvent.html

Tody, D., & Plante, R. (2004) Simple Image Access Specification, Version 1.00. Retrieved from the World Wide Web, September 16, 2006: http://www.ivoa.net/ Documents/latest/SIA.html

Wells, D. C., Greisen, E. W., & Harten, R. H. (1981) FITS—A Flexible Image Transport System. *A&A Supp. Ser. 44*, 363-370.

World Wide Web Consortium (2004) World Wide Web Consortium Process Document. Retrieved from the World Wide Web, September 16, 2006: http://www.w3.org/2004/02/Process-20040205/