A Case Study of The National Virtual Observatory as a Digital Curation Project

Ramona Broussard, Rebecca Holte, Bridget Jones, and Emily Vinson

University of Texas at Austin

Abstract

Astronomers need access to relevant data in order to conduct research and discover knowedge. In the astronomy field a single data set can have many uses. If astronomy data is widely available, that availability will facilitate advances in astronomy research. The National Virtual Observatory (NVO) has developed a framework for data registry, creating online access to distributed astronomical data sets. The extensive range of tools provided by the NVO, give researchers access to data sets (such as images, spectra, catalogs, light curves, and records of transient events), providing searching, visualization and analytical capabilities. The NVO provides a means to publish and access data that is stored elsewhere. The strong standards and technical structure of the NVO project create a foundation for the online data collection. In addition to accomplishing its creators' original intent of usefulness to the astronomy research community, the NVO provides a model for other online digital curation projects to follow in their pursuit of providing access and storage for long-lived, authentic, and extensible data collections.

Keywords: National Virtual Obervatory, Virtual Observatory, Virtual Sky, Data Curation, Digital Data Collection

The National Virtual Observatory (NVO) is an example of a digital curation and cyberinfrastructure project that has been highly successful, especially when compared to similar curated projects that are now defunct or empty. Astronomy data and metadata has been relatively standardized compared to subject areas whose data sets may include books, tables, photographs, oral histories, and GIS information. Studying the methodologies of the Virtual Observatory framework will prove useful for other online data collections and portals, localized and distributed, scientific and humanist.

The primary goal of the NVO is interoperability among astrophysical data sets; the project has culminated in a framework and portal allowing remote access to instrumental data sets obtained from land- and space-based telescopes and space missions. The normalization of astrophysical data, and developed metadata standards, has allowed the coordinated development of an array of tools allowing the discovery, integration, comparison, and analysis of the distributed data that has not been possible in the past. While the NVO is not a repository for data sets, it lays the groundwork for the publication of data sets, preservation of queries, and new forms of scholarly communication and data curation (Choudhury, 2008).

History and Growth

"Astronomy and Astrophysics in the New Millennium"

"As the new millennium begins, astronomy faces a revolution in data collection, storage, analysis, and interpretation of large data sets." (National Research Council [NRC], 2001a, p. 132).

In 1997 a group of leading astronomers convened to discuss a new decadal survey of astrophysics and astronomy. As a result, the Astronomy and Astrophysics Survey Committee

(AASC) was created with the charge of completing an assessment of the field and recommending programs and priorities for 2000 - 2010 (NRC, 2001a, p. xv). The AASC completed its final recommendations in 1999 with the involvement of many agencies. Both the main report and panel reports were published in 2001 by the National Research Council under the title *Astronomy and Astrophysics in the New Millennium*.

The NVO was identified as the top priority under Small Initiatives with an estimated Federal cost of $60M over 2000-2010 (NRC, 2001a, p. 10). A program spanning both ground- and space-based science communities, it proposed to create a "virtual sky" and allow greater access to large astronomy data sets for both specialists and the general public. It aspired to provide an "unparalleled opportunity for education and discovery" (NRC, 2001a, p. 14). Notably, many high-priority initiatives were anticipated to make data available through the NVO, such as the Large-aperture Synoptic Survey Telescope (LSST), as well as existing projects, such as the Two Micron All Sky Survey and the Sloan Digital Sky Survey (NRC, 2001a, p. 132).

The AASC foresaw that the NVO could help to increase science literacy. "Trillions of bits of information" were anticipated to be added to the NVO on a daily basis; the NVO would virtually link the major data assets, utilizing computers distributed across the country in a high-speed network. To augment the capabilities of this unified collection, the NVO needed data standards, data mining tools, services and tools for advanced analysis, and linkages between the research data and education system. (NRC, 2001a, p. 132).

"Toward a National Virtual Observatory"

In 2001, coordinated efforts began with the assistance of a Small Grant for Exploratory Research, AST-SGER-9876645 (National Science Foundation [NSF], 2001b, p. 86), awarded to Alexander Szalay of Johns Hopkins University, supporting him in the identification of key issues, enabling community discussions, and a variety of publications.

The NVO Interim Steering Committee endorsed the white paper "Toward a National Virtual Observatory" in 2000. The paper identified the scientific opportunities, technical challenges, and implementation strategy for realizing the goals of the proposed NVO. "Technology-enabled but science-driven", the NVO would serve as "an engine of discovery for astronomy" and "a venue for educational outreach" (National Virtual Observatory [NVO] Interim Steering Committee, 2000, p.2).

The NVO took advantage of advances in computational speed, storage media capacity and detector technology. New generational surveys span a range of wavelengths, and new software tools may be developed to fully analyze the petabyte databases that are subsequently generated, enabling new qualitative discoveries. Historic examples include the combining of optical and radio wavelength images which led to the discovery of quasars; repeated imaging of regions of the sky led to the discover of transient phenomena such as supernovae (NVO Interim Steering Committee, 2000, p. 7).

It was clear that implementation would face significant challenges, specifically in the incorporation of existing data archiving efforts and the development of new capabilities and structures. The AASC recognized that a community-driven "Science Reference Mission" (SRM)

would be a necessary tool in guiding NVO development decisions. They proposed a structured process of meetings, beginning with a community-wide workshop in Pasadena in June 2000. Working groups were charged with detailing major programs, explaining the scientific merit, and articulating the flow from scientific needs to access and tools requirements.

The AASC also envisioned that close collaboration with the computer science community and early functionality were essential to the project's success. To ensure early functionality, they enacted a four-phase process, and NVO management activities were to maximize community participation. The levels of activity and funding were meant to provide a usable and well-documented infrastructure, encourage science-driven software tool development and disperse the funds to appropriate peer-reviewed projects. Initially, the highest priority was to build archive infrastructure and access protocols (NVO Interim Steering Committee, 2000, p. 19), but the activity levels were expected to find a balance over time. The goals of the four phases would ideally be implemented within five years of inception (NVO Interim Steering Committee, 2000, p.20-21):

- Phase 0 (prior to NVO start): Create conceptual design of the NVO; begin implementation activities at select centers.
- Phase 1: Establish integrated services for data discovery, delivery and comparison.
- Phase 2: Establish initial cross-correlation capabilities and begin full-scale operations.
- Phase 3: Establish fully functional baseline NVO and enable full-scale cross-correlations capabilities.

These preliminary efforts culminated in the award of a 2001 NSF Information Technology Research (ITR) grant supporting creation of the NVO framework (National Science

Foundation [NSF], 2001). Meanwhile, "NASA continued to support the creation and maintenance of archives from space astrophyiscs missions and their distributed data systems through the NASA centers" (NSF, 2008).

The NSF grant application reiterates the enabling of new scientific methodologies and introduces the utilization of the NVO as a large-scale prototype of the "semantic web" (NSF, 2001, p. 3) by federating widely distributed, heterogenous datasets. The application also discusses the wide-spread problem of how to publish scientific data in addition to papers. The NVO is proposed as a facility for data publication, providing "semantically-rich" information and recognizing the different roles of author, publisher or curator, and reader. It is emphasized that standardization would only be used for locating and federating the data sources, and the end user would always be able to obtain data in its original format (NSF, 2001, p. 4).

In order to realize the above goals, an NVO Testbed was proposed to provide a platform for testing tools, standards and prototypes with real data from select institutions such as Caltech, and JHU. Key activities include software installation and testing; resource allocation and scheduling procedures across computing and storage resources; development documentation and user support; problem identification and recommendations; and the facilitation of data collections movement among the participating sites. Project components would be implemented incrementally, providing user access to new capabilities at regular intervals.

A management plan was also outlined in the 2001 funding application, which proved challenging due to the project's distributed, multidisciplinary nature. The 17-organization collaboration is co-led by Dr. Alexander Szalay at Johns Hopkins University (JHU) and Dr. Roy Williams at California Institute of Technology (Caltech) as the grant's Principal Investigators (at

the time of the grant, Dr. Paul Messina was the Caltech PI). Szalay also designed the architecture for the Sloan Digital Sky Survey. Robert Hanish, of the Space Telescope Science Institute, was appointed project manager and later began to serve as chair of the International VO Alliance (IVOA). Notably, their expertise spans astronomy and computer science.

A contractual relationship exists between the NSF and JHU. Their management approach relies on clearly-defined work packages, deliverables, and schedules with designated check-points. The executive committee maintains full authority to terminate the participation of unproductive groups. Senior staff, collaborators and international liaison were established at the time of the funding application.

Core responsibilities in the NVO are as follows (NSF, 2001, p. 16):

- NVO Executive Committee: approve or reject changes to technical development plan, review progress, allocate (or re-allocate) funds. The Principal Investigators co-chair.
- Program Coordinator: ensure that development tasks meet science goals and that incremental capabilities are delivered to the end-user.
- Project Manager: monitor schedules, budgets and progress of task teams.
- System Architect: coordinate overall system engineering and design functions, including interfaces between elements.
- Technical Working Group: provide coordination and collaboration to the Development Task Teams; it is head by the Project Manager and includes leads of each Team.
- Development Task Teams: Portals and Workbenches, Metadata Standards, Grid Services and Testbed, Data Models, Resource Layer and Data Access, and Data and Services.

- Science Working Group: define, develop, deploy and test science prototypes; it is chaired by the Program Coordinator and includes application scientists, the Project Manager and the System Architect.
- Education and Outreach Coordinator: ensure NVO efforts benefit education and the public; designated team members ensure coordination with related international efforts.

At the same time as these separate efforts, and following AASC recommendations in 2000, NASA and the NSF established the Science Definition Team (NVO SDT), to further refine the NVO project. Between 2001 and 2002, the NVO SDT prepared a report that incorporated the AASC report, the 2000 draft white paper, the NSF Information Technology Research Grant proposal, community input, and proceedings from NVO development workshops. The report addressed the following key questions:

- Project definition,
- Implementation phases,
- Estimated costs,
- Management structure,
- International relationships, and
- New science and societal benefits.

The NVO SDT distinguished the NVO from other archives and similar projects based on its wide range of abilities; the NVO can provide cross-archive searches, correlations, and analysis, as well as computational and data management services. The NVO is at an advantage over past projects due to industry-wide advancements in distributed computation and storage capabilities, over a decade of experience in data management, information services, and archival

infrastructure (the SDT cites NASA mission data sets as an example). (National Virtual Observatory Science Definition Team [NVO SDT], 2002, p. 9).

Growth

"We will know that the NVO is a success when it is used daily by thousands of astronomers, educators and members of the general public - and that it is taken for granted, just like the most successful web-based information services today (NVO SDT, 2002, p. 12)"

2001 – 2002 (year one of NVO).

The initial $14.1 million award for the purpose of establishing the National Virtual Observatory was awarded to Johns Hopkins University and Alexander Szalay in September 2001 (NSF, 2001).

Phase one of the NVO was planned for 2002 - 2003, this included the conceptual design and early development work (NVO SDT, 2002, p.42). During this time the NVO team tested interfaces and protocols such as the VOTable standard, MONTAGE (Mosaics On the TeraGrid Express), ROME (Request Object Management Environment) and data modeling. By the end of 2001, VOTable was proposed as the XML standard for representing astronomical data tables (Ochsenbein, n.d.) and in 2002 it was accepted internationally (Williams, 2002; Williams et al., 2002). VOTable was rapidly embraced by astronomers and in 2002 VOTable was implemented in 50 cone search services (NVO, 2002b, p. 1).

In addition to the cone searches, the NVO experimented with Simple Image Access Protocol (NVO, 2003b). Registries for the NVO were designed so that new services would have

their metadata cached by the NVO, “When an implementing service registers itself, the central registry will send it a metadata query and cache the results" (NVO, 2002c, p. 8). Two large sky surveys, the Sloan Digital Sky Survey (SDSS) and 2MASS survey, were involved in early testing for registries and SIAP. Universal Content Descriptors (UCDs) were assigned to attributes in the SDSS (NVO, 2002c, p. 11) and testing the UCDs helped fill in the gaps of the UCD hierarchy (p. 11).

At the end of 2002, the NVO Advisory Committee met and wrote a report praising the NVO for their commitment to standards but warned that the semantic web is volatile (NVO Advisory Committee, 2002, p.7). The committee suggested that the NVO develop a formal schedule and management plan to track their progress while they develop the framework for the NVO (p. 1). The NVO responded with a management plan, which documented the organizations and individuals involved with the NVO and their anticipated contribution.

2003 (year two).

After the successful partnership of the NVO and IVOA to develop VOTable, the NVO went into 2003 hoping that the same model of cooperation could be used to make a standard Simple Image Access Protocol (SIAP) (NVO, 2002c, p. 3). At this point, the XML schema VOResource was implemented for the SIAP-side of cone searching, “We have applied this model to describe Cone Search and Simple Image Access services, including their inputs and the columns found in their output VOTables" (NVO, 2003b, p. 8). Using the VOResource schema and VOTables, the NVO was able to test out the inventory service.

The development of the STScI registry raised questions for what exactly would be registered with the NVO. In 2003 it was decided that the registered resources should be "curatable", meaning that some outside entity would be responsible for the resource. The issue of granularity was also discussed in regard to what exactly a resource is (NVO, 2003b, p. 8). The ultimate definition of what a resource is remains somewhat nebulous:

> A resource is a general term referring to a VO element that can be described in terms of who curates or maintains it and which can be given a name and a unique identifier. Just about anything can be a resource: it can be an abstract idea, such as sky coverage or an instrumental setup, or it can be fairly concrete, like an organization or a data collection. (Benson et al., 2009)

The Simple Image Access Protocol was developed to prototype how to access astronomical data in a first year demonstration, it was not initially intended for widespread use (NVO, 2002d, Overview Section). However, SIAP gained attention internationally and in 2003 SIAP was useful for populating the STScI registry. By the end of 2003, the registry of SIAP compliant services reached 100. Many institutions with sky surveys made their services SIAP compliant, the Infrared Science Archive (IRSA) made one of the largest contributions to this effort, the services they registered provide access to multiple sky surveys at Caltech-Infrared Processing and Analysis Center (NVO, 2003b, p. 30).

Also in 2003, the High Energy Astrophysics Science Archive Research Center (HEASARC) developed a web interface to search the Space Telescope Science Institute (STScI)

registry for SIAP compliant cone search services (NVO, 2003b, p.9). The Data Inventory Service portal was the first ever end-user service officially released by the NVO (NVO, 2003b, p. 9).

2004 (year three).

In 2004 some final infrastructure work was done on the registries in preparation for making them public (NVO, 2004, p. 55). The improvements to the NVO registries in 2004 included:

- Boolean keyword search capability,
- The establishment of mirror sites, including an STScI mirror site,
- Testing of the VOResource 0.10, and
- The development of Carnivore as a resource registry service.

After populating the NVO registries for two years, curation became a topic of interest for the NVO; they began encouraging the improvement of metadata in the registries. One approach for metadata was a verification flag. The verification would be either automated or human and applications could flag the filters to find compliant services. The SIA service registries began using verification tests that were already being used for cone search services (NVO, 2004, p. 16).

Also in 2004, JHU began development on OpenSkyQuery, which is a protocol and interface used to search for registered SkyNodes (NVO, 2004, p. 15).

2005 (year four).

In 2005 the NVO continued discussing issues of data preservation and data curation. In particular, the NVO started to explore how to curate data that was used for peer-reviewed

literature (NVO, 2005a, p. 4). Caltech started work on a curation workflow in 2005, "starting with automated checking tools, then human checks, and a way in which corrections can be made by interacting with the original publisher of the records" (p. 40).

To ensure authenticity of metadata, a "tiger team" from the NVO began reviewing all of the metadata in the resource registries for consistency (p. 11). They started with the primary NVO partners and reported feedback to publishers about what the problem areas were. In some instances, the NVO worked directly with publishers and many of the problem spots were used to improve the instructions given to resource publishers (p. 12). To assist in metadata consistency the registry element ResourceValidationLevel was added to the resource metadata schema for registry curators to use to rate records based on levels of conformance to metadata standards (p. 10)

As of 2005 there were nearly 20,000 resources registered in the VO (p. 30). At this time the number of registries was increasing and the proposal to develop a "registry of registries" was brought up, it would be maintained by the IVOA and would identify the purpose of each registry (p. 4).

2006 (year five).

This was the final year for the NVO's infrastructure development. Nearly all the services that currently exist on the site were first deployed on or before January 2006 (NVO, 2006a, p. 1). The NVO teams anticipated that future In addition to the cone searches, the NVO experimented with Simple Image Access Protocol (NVO, 2003b). Development of the NVO

would be geared toward building tools for general audiences and work on Education and Public Outreach (EPO) (p. 2).

VOEvent software was a major release for the NVO in 2006, VOEvent is "a standard for communicating reports of transient astrophysical events, with the intention that automated systems will interpret such messages, then make a follow-up observation, then report the follow-up as further VOEvents" (NVO, 2006a, p.35). To be effective, the VOEvent began integration with the registries, it would need to be defined by publishers of VOEvents and repositories of VOEvents (p. 12.)

In this last year of NVO development, the team tried to finalize the registries and establish "registry curation practices". They decided to start describing services in terms of the service capabilities, "this allows one to register support for several standard and non-standard interfaces to the same underlying collection in one registry description...The clean separation between collections and services will make it easier for users to find what they are looking for" (NVO, 2006a, p.11). The NVO formally addressed granularity in the registries, deciding that information on table columns would be optional for registering resources. When registering a resource full records are uploaded, but it is not required to make all of the columns searchable by describing them in the search registry. (NVO, 2006a, p.11).

At this time an automated validation service was being developed, the service would allow pushlishers to submit their metadata and get a response about metadata inconsistencies. The serivce was tested on HEASARC and STScI and it did improve metadata compliance (NVO, 2006a, p.13).

2007 (year six).

2007 was an extended sixth year of development for the NVO, but there was not enough funding to continue development at the rate of previous years; the work on EPO and user-friendly interfaces was stagnant. To continue the project the NVO applied for and received funding from NASA and the NSF (NVO, 2007, p.1).

The basic architecture for the NVO was deployed by this time and the NVO began to shift their focus from infrastructure to more general issues of sustainable stewardship. The Data Curation and Preservation Interest Group was formed to ensure data authenticity (NVO, 2007, p. 7). The NVO continued to work directly with data providers to revise non-compliant metadata in the registries and a revision of the process for how to register services was underway (NVO, 2007, p. 8). One other effort toward compliance was the continued production of a registry validation service that would validate services before publishing them in the registries (NVO, 2007 p. 10).

2008 & 2009 (years seven & eight).

By 2008 there were several NVO services that provided tabular data as query responses. For that reason the NVO has been working on the Tabular Access Protocol (TAP) to standardize how these tables are accessed (NVO, 2008a, p. 9). Other progress made in the last two years included the completion of The Registry of Registries (International Virtual Observatory Alliance, n.d.) and an update to the NVO website. The current public NVO website (the NVO data discovery portal) was released February 2009 and the registry was renamed the "directory" (NVO, 2009a, p. 9).

"The Role of the Virtual Observatory in the Next Decade"

"With the new NVO portal deployed and the project operating on remaining no-cost extension funding, our current efforts in the area of Registries are concentrated primarily on maintenance of current services and participation in IVOA standards efforts." (NVO, 2009a, p. 4).

"The Role of the Virtual Observatory in the Next Decade", released by the US Virtual Observatory Consortium (2009c), reiterates that the NVO mandate did not include a "user ready" entity. Rather, the National Virtual Observatory was the first initiative with a goal to establish the infrastructure necessary for a functioning Virtual Observatory (the world-wide federation of data and computational resources). Tools and applications were created, but with a core audience of specialists and technical experts, who would in turn test the new systems, becoming expert users as well as NVO advocates.

The second major initiative in realizing the VO concept, is the establishment of a fully functional, operational entity: the Virtual Astronomical Observatory (VAO). Like the NVO, it will be distributed, responsive to user needs and demands, and committed to the implementation of new technologies. The emphasis furthers the original NVO goals, in the facilitation of research for all astronomers, not just the technical experts. It will continue to build upon the framework, infrastructure, and tools created by the NVO project and the NASA astrophysical data centers, and operate within the context of the International Virtual Observatory Alliance (IVOA)(NSF, 2008).

The VAO initiative creates an environment where researchers have access to diverse data sets (such as images, spectra, catalogs, light curves, records of transient events, etc), are able to browse, visualize, and analyze using NVO-created or "VAO-aware" tools. The VAO is not a central repository, nor a system of closed, proprietary software. Rather, it is a means to publish and access data. The VAO relies on data providers to maintain accuracy as well as the peer-review process associated with the published literature, but will confirm the accuracy of the description, to aid discovery and utilization of the data (US Virtual Observatory Consortium, 2009, pp 4-5).

In 2008, the NSF released the program solicitation for the Management and Operation of the Virtual Astronomical Observatory, offering joint funding with NASA, in the amount of nearly $6M annually for the five-year cycle ending with FY2012 (NSF, 2008). In February 2009, the NSF (and NASA) declared intent to fund the proposal. NASA has begun its funding share, but the NSF has yet to fully implement its portion of the award (Robert Hanish, personal communication, October 1, 2009).

See Appendix A for a list of milestones by year.

Literature Review

There are a wide range of sources available that discuss astronomy collections, grid computing and the NVO itself. For example, the National Virtual Observatory website contains a great deal of project documentation (www.us-vo.org). In addition to the NVO document repository, authors representing astronomy organizations have written in many journals and magazines including Computer World, the journal for the Association for Computing Machinery,

and the National Academies Press. Most of the literature to date is lacking in serious critiques of the project, and has a greater focus on description of tools or overall goals of the NVO.

It is clear that the astronomy community widely knows about and uses the NVO. Perhaps because of this broad acceptance of an international effort, which is in many ways ahead of similar digital curation projects in a variety of subject areas, there is a shortage of case studies or contemplative evaluation reports in the documentation. However, from the ample collection of documentation and description of the NVO project, several themes are apparent.

Open access and federation of information in the NVO

Looked at simply, a Virtual Observatory is a portal for the search and analysis of data sets from a variety of distributed sources. In addition, the Virtual Observatory framework defines the cyberinfrastructure for the astronomical community. The NVO is an instance of a Virtual Observatory; standardization of data sets allows it to successfuly function as a portal.

The NVO is unique from many other digital curation projects in that it aggregates access to data from disparate sources. In their 2001 paper, "The World-Wide Telescope", Jim Gray and Alexander Szalay describe how the National Virtual Observatory brings together data from multiple sources to create an open-source website that facilitates teaching and discovery. The NVO provides access to data from the Hubble Space Telescope, the Chandra X-Ray Observatory, the Sloan Digital Sky Survey, and other compliant resources (Gray, p. 2037). Gray and Szalay explain that the project also reaches an international audience.

According to Szalay in "The National Virtual Observatory", the NVO is intended for a wide audience, including astronomers, students, teachers, and the general public (p. 3). Diversity

means that the NVO must be accessible to those audiences online, and must have tools that extend beyond specialty tools used by astronomers. Anyone can create an account to access the tools provided on the NVO website, and gain access to the combined data sets there.

The role of the NVO as a federated source for astronomy databases is well-documented. "The Management and Operation of the Virtual Astronomical Observatory" contains the application guidelines for a NSF/NASA funded grant to manage the Virtual Astronomical Observatory (VAO), the next phase of the NVO. The VAO will "serve to link a multitude of astronomical data sets into an integrated system that allows automated search and analysis among all cataloged objects" (p. 1). The expectation of the grant recipient's implementation would be to make use of the standards established by NVO. The Introduction section of the document (p. 4) provides a good background description of the NVO. The bulk of the document describes the grant application process.

Krumenaker's article describing the relationship of astronomy and the internet synthesizes astronomy resources, many of which are brought together via the NVO. Describing astronomy as the "most intangible and least experimental science," the author describes how astrophysics as a science is moving out of the observatory and in to the office. The astronomy field was one of the first to rely heavily on computation, and to move on to the web – as a place for journals, searchable databases and data sharing. In large part, the article is dedicated to a list of particular examples of astronomy repositories online.

The necessity of scalable metadata

In order to aggregate this data from differing projects and telescopes, the NVO needed a strategy for creating metadata that was robust and could also scale across collections. Raymond L. Plante writes, "Naturally, expanding capabilities will drive the need for increasingly detailed information to be exchanged in the form of metadata." (2009) There is an international standard for astronomy data known as FITS (Flexible Image Transport System). The wide-spread file wrapper standard is an essential underpinning for the NVO project, and likely one reason for the success of the NVO. A scalable metadata schema is one that can expand with the addition of more detailed metadata without increasing the cost of supporting the metadata. According to Plante, the VO structure uses metadata both to describe resources and exchange metadata between programs, humans, and services. The metadata provided must support all of those activities.

"Space-Time Coordinate Metadata for the Virtual Observatory" outlines the International Virtual Observatory Alliance's (IVOA) recommendations for metadata standards in virtual observatories. It is significant that the IVOA "now comprises 17 VO projects from Armenia, Australia, Brazil, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the United Kingdom, and the United States", thus these standards are not limited to the NVO, but apply to virtual observatories all over the world. Because of this international effort, the standards developed reflect an effort to accommodate different equatorial systems, historic data, as well as readings taken from space. The metadata developed is primarily concerned with the meticulous specifications of coordinate axes, which include: Space (including

its time-derivatives, i.e., velocities); Time; Spectral (frequency, wavelength, energy); and Redshift (Doppler velocity).

The portion of this document concerned with the actual metadata standards is highly technical, and not geared toward the lay person, however suffice it to say that the IVOA has created very specific methods of metadata production, as well as an XML Schema its mark up.

The bulk of the 2004 NVO meeting minutes document does not deal with the NVO, only a brief paragraph discusses NVO. Of note is a reference to the fact that many datasets were already (in 2004) NVO-compliant in terms of protocols and standards. Widespread compliance with global standards may be one reason for the success to date of the NVO project.

Grid computing and technology tools

In a paper on computational challenges in astronomy, Babu and Djorgovski discuss the challenge of processing terabytes or more of data in astronomy. The astronomy community contains multiple large and multi-faceted data sets; the NVO emerged to standardize the structure of the databases that hold the data sets and federate tools to perform search and calculations on those data sets. The authors write that the NVO will strengthen communication between astronomy communities and strengthen statistical analysis of astronomy data sets by providing a computational tool for the community. (p. 4)

In "Grid-Based Galaxy Morphology Analysis for the National Virtual Observatory", the authors describe how the NVO provides a portal to access distributed databases using a grid. This interface standardizes user access to a variety of astronomy data sets with diverse size, interface design, and content. The grid-based nature of the VO project does require

standardization of metadata and formatting, which in turn allows programmers to provide processing algorithms for data processing. Programs to create galaxy morphologies using the NVO's grid is one way these data sets have been brought together by researchers.

In the chapter, "Deja vu All Over Again", the authors describe using tools in the NVO for their project to revisit data from snapshots of radio galaxies taken in 1989. In this project, the authors were able to use the NVO structure, including the metadata standards and computing resources, to create a simple web service that would perform new calculations on a set of data almost 15 years old. They also used standard VO tools, for example a cone search, to check their calculations. (p. 2)

Virtual Sky and Montage mosaicing services were early examples of astronomical grid computing. These services were the first to federate sky surveys such as the Sloan Digital Sky Survey, Hubble Deep Field, Two Micron All Sky Survey and Digitized Second Palomar Observatory Sky Survey. Once federated, users are able to sample images from multiple wavelength telescopes and overlay the images. Stacking images can help astronomers detect faint objects and federating multi-wavelength images provides astronomers with multiple perspectives of the same image. The author, Roy Williams, was also a member of the intial VOTable team; in Grids and the Virtual Observatory he explains how VOTable was designed for the datasets found throughout the astronomy grid.

Remote access in the field of astronomy

In the June 2000 article, "Mining the Digital Skies", the author discusses the emerging relationship between databases and telescopes: namely that large databases allow astronomers to

make observations without sitting in front of a telescope, but rather requesting particular data from observatories via “service observing.” One benefit of these large databases is the potential of using tools to discover patterns in the large data sets. At the time the article was written, ASTROVIRTEL (Accessing Astronomical Archives as Virtual Telescopes, 2000-2003) was the primary example of a successful astronomy data collection set, however the National Research Council had allocated $60 million the month previous to establish the NVO.

In a 2001 *Computerworld* interview, Neil de Grasse Tyson, Director of the Hayden Planetarium and member of the NVO steering committee, discusses the future of the NVO. Of particular interest is de Grasse Tyson’s view that the physical location of the observatory is irrelevant because of the Internet and AI tools to mine data. He also considers the importance of data quality and metadata standards in creating a tool valuable to all NVO members. Finally, de Grasse Tyson discusses the democratizing effect the NVO will have: no longer will the big discoveries be limited to members of “rich institutions.”

The authors of "Relying on Electronic Journals: Reading Patterns of Astronomers" discuss information seeking behavior of astronomers. The article is entirely devoted to the question of how astronomers gather published information about their research interests - print journals, electronic journals, or e-prints. This study is distinctly about the preferences of astronomers undertaking research, such as literature reviews, not about the use of data sets.

Transparency in planning and reporting

The functional origins of the NVO initiative can be found in the decadal survey report titled "Astronomy and Astrophysics in the New Millennium", where the NVO was identified as a

priority for the next decade, 2000 - 2010. From that point on, the NVO has diligently made available an array of planning and operational documents, not only due to its public funding status, but also to create a clear framework and to inform the development of similar projects. Available via the NVO website, http://www.us-vo.org/pubs/index.cfm, documents include white papers, quarterly and annual reports, science and technical reports, and education and outreach reports. The repository includes periodic Advisory Committee Reports, NVO Summer School summaries and surveys, and reports on a variety of released services. The direction of the NVO, accomplishments, and lessons learned may be traced through the documents provided here.

Project Content

The National Virtual Observatory project was built for astronomical data. They accept data from any telescope, not dependent on location, but on registration and relevance of data; the data should add to the astronomy field. Because of their commitment to community needs and data preservation, the NVO has done an excellent job of creating metadata standards and tools for their collected data. This has led to the voluntary submission usable astronomy data, and seems to have prevented any great need for an active collection development policy. Support of the community through education has also helped the visibility of the NVO.

Collection Principles

The NVO is still actively accepting submissions from observatories. Some organizations that the NVO collects data from are static, that is, they have submitted the final piece of data that they will submit. One example is the Sloan Digital Sky Survey (also referred to as SDSS); SDSS

completed their survey in 2008. Other organizations such as The Two Micron Sky Survey (or 2MASS) are still collecting and providing data to the NVO portal.

According to NISO, "A digital collection consists of digital objects that are selected and organized to facilitate their discovery, access, and use. Objects, metadata, and the user interface together create the user experience of a collection." (NISO, 2009) The NVO has selected objects by restricting the collection to data relevant to astronomy, and remains committed to discovery, access, and use; in fact, according to the NVO 2005 project update, the NVO is dedicated to the following activities (pp. 2-3):

- Managing large-scale computation and data storage,
- Scalability,
- Easy conversion across collections,
- Quality via registration, and
- Tools and technical projects.

The National Science Foundation website's "Cyberinfrastructure Special Report" explains that digital libraries are, "Meaningful collections from all facets of society must be compiled, structured and preserved. Increasing computational power and network bandwidth must be applied to make these collections accessible, usable and interoperable, and interfaces to these collections must be designed to be clear, flexible and scalable." (NSF, 2008a, ¶. 2)

The goals of the NVO fit closely with the NSF description of the function of digital digital libraries by compiling and structuring astronomy data, providing computational power to

the astronomy community, providing general access to these collections, focusing on scalability, and developing tools for the website that interact with the collection data.

Ingesting data.

Anyone that has data of interest to the astronomical community may upload data to the NVO. Data must be accompanied by metadata, and should be formatted in XML. This process is also known as publishing data. It is also possible to ingest (or publish) services to the NVO, meaning those programs or applications that are written to work with NVO data. Submitters can decide how much exposure they want their data to have during the ingest process.

It is important to remember that the NVO is not a data repository. Although anyone, no matter how small their data set, can ingest data, they will then have to host it. If that is not viable, the NVO suggests depositing the data at an open repository that is VO compliant. In addition, according to the NVO's "How to Publish" web page, "Open repositories are more than just a web site that will host your data. They take responsibility for long-term curation of the data" (2008c).

The most basic level of the submission process is registering with the NVO. This will provide the NVO and related applications with important administrative metadata about the organizations that are publishing data. This is also the point at which each collection is assigned a unique identifier within the NVO structure. Once a data set is registered, it is possible to move to the next layer of exposure. This layer depoys the data to standard NVO services, such as cone searched. More about these services can be found in the tools section of this document.

To interact with programs of the NVO, metadata should be XML formatted; this is simple for data providers without their own XML standards and procedures because the NVO provides a web front end for entering metadata in the right format.

Decision making in the NVO.

The NVO collects and aggregates data from land- and space-based instruments, with an emphasis on extracting quality metadata. Alexander Szalay is the principal investigator and project director for the NVO. Since the NVO is funded by the National Science Foundation, Szalay is responsible for reporting to them. In addition to the team of project managers and architects, the NVO has an external advisory committee to ensure that community needs are being met. (NVO 2004, pp. 2-4)

Object Characteristics

Astronomy data ingested into the NVO is data collected from telescopes either on the ground, or in space. Astronomy data objects are records of transient or stationary astronomical events, these are usually in the form of images or catalogs. Data objects in the NVO have a table of metadata that includes fields for metadata such as description, title, publisher, waveband, identifier, short name, and categories (output type).

Data sets are not checked for integrity over time via a method such as a checksum verification. The NVO is a portal to search, not a preservation project, so the NVO requires only registration and performs quality assurance on the data sets that they allow in the collection by assuring that the metadata follows standards. The project includes ample documentation, publicly available on the website. However, quality of metadata does not necessarily disclude

tampering or "spoofing" of data by the original scientist. It should be noted that all digital objects come from a limited number of telescopes that are "trusted." In an email to the authors, Robert Hanisch mentioned that the astronomy community is relatively small; in this case, prestige of the data provider serves as an authenticity check (personal communication, November 13 2009).

The content management system is unique to the NVO. According to Sayeed Choudhury in his article, "The Virtual Observatory Meets the Library", "The essence of the Virtual Observatory is interoperability. Data discovery, data access, and database queries are enabled by metadata standards. A primary goal of the Virtual Observatory is to provide integrated access to archival data and derived data products: catalogs, tables, and highly processed images, spectra, and time series." (¶ 7, 2008)

The goal of the NVO is not to collect or store data, rather it is to provide access to the already existing astronomy data from a variety of repositories. Currently, Choudhury writes that the NVO scope has, "...not included long-term data curation, focusing instead on data location and data access standards and protocols." (¶ 8) It is perhaps this narrow focus that has enabled the NVO to develop strong ties with the community and to provide robust metadata standards.

According to Choudhury, in the case of the NVO, data are publications. (¶ 9) Data in the NVO is level three and four data, which means that that data can be cited in traditional journal publications. This is very important to the data sharing issue, because scientists are more willing to share data if they have an established avenue for credit or prestige. This may be another factor in the success of the NVO. They have had to change sequence of publication; publishing now comes before analysis. In the past publication has been the last step, but this is no longer the best

way to operate. Databases such as the NVO are becoming more a part of scholarly communication.

Metadata

NVO metadata.

Over the course of several years, the NVO Metadata Working Group, as well as the IVOA Web and Grid Services Working Group, sought to create metadata standards for use by members of the NVO, as well as international VO members. The highly specialized characteristics of the astronomical data collected by the NVO requires very specific metadata standards. The NVO Working Group settled on a hierarchical system to manage the VO metadata. This system is broken into a top level, which calls for a minimum of information, and lower levels that have more involved metadata requirements, "allowing for the description of query syntax, access protocols, and usage policies." (NVO Metadata Working Group, 2002, p. 1)

The definitions for the following terms list are taken from the NVO Metadata Working Group documents as well as the IVOA Support documents:

Service: "any VO element that can be invoked by the user to perform some action" (NVO Metadata Working Group, 2002, p. 1).

Service metadata: "describe the service's interface as well as information that aids in effective use of the service…represented as an XML document" (Web and Grid Services Working Group, 2009, p. 4).

Query service: “supports query/response protocol" (NVO Metadata Working Group, 2002, p. 1).

Registry: “a query service for which the response is a structured description of other services" (NVO Metadata Working Group, 2002, p. 1).

Resources: a collection of one or more services, or other resources, that share common metadata characteristics (e.g., Publisher, Creator, Contributor, Identifier, Contact, Type, Facility).” (NVO Metadata Working Group, 2002, p. 1).

Curation metadata: “describe who supports the resources and what its purpose is" (NVO Metadata Working Group, 2002, p. 2).

Content metadata: “describe what kind of information is available" (NVO Metadata Working Group, 2002, p. 2).

Resource metadata: “encompass information that is simply known to users; constitute a ‘yellow pages’ of astronomical information" (NVO Metadata Working Group, 2002, p. 2) and are human readable.

Capability metadata: “interface that provides the service metadata in the form of a list of Capability descriptions, XML element that 1) states that the dervice provides a particular; IVOA standard function; 2) lists the interfaces for invoking that function; 3) records any details of the implementation of the function that are not fixed in the standard for that function." (Web and Grid Services Working Group, 2009, p. 3)

Capability list: "describes the service to software... has two uses: 1) may be read by client application to find out how to drive the service; 2) may be read by the registry to compile the registry entry for the service" (Web and Grid Services Working Group, 2009, p. 4)

Availability metadata: indicates whether the service is operable and the reliability of the service for extended and scheduled requests; represented as an XML document, contains child elements: available, upSince, downAt, backAt, and none. All elements except for available are optional. (Web and Grid Services Working Group, 2009, p. 5)

Table metadata: represented in XML, as defined in the VODataService standards (Rixon, 2009, p. 5) .

In addition to the above metadata, which describes curation, content, and accessibility, NVO metadata also describes the physical location of the object being described, as well as the location of the observatory capturing the image and the time and type of the capture. These metadata include the coordinate system, coordinate values and coordinate areas or ranges. This is only a small sample of the metadata collected that relates to location, other data collected includes pixel, astronomical and spatial coordinates, time and space frames, and region metadata. For each of these examples are many specific components. For example, the Coordinate has six components: CoordName, CoordValue, CoordError, CoordResolution, CoordSize, and CoordPixS (NVO Metadata Working Group, 2007, 4.4.1).

Some fields of the NVO metadata schema require controlled vocabulary. The Types field is an example of an NVO field with a controlled vocabulary. This field, an element in of Content Metadata, describes what aggregate the data is gathered in, thus the NVO working group

recommends that users limit themselves to the controlled vocabulary: archive, survey, catalog, bibliography, journal, library (NVO Metadata Working Group, 2002, p. 3).

Materials are marked-up in XML; VOTable is an XML standard based on Dublin Core (NVO, 2003a). Even though VOTable is based on Dublin Core, there are many VOTable elements that did not originate from Dublin Core (p. 4). Coverage is one particular element that has many additional subset elements because coverage element in Dublin Core is too broad.

Flexible Image Transport System (FITS) is the most common image and data file format for astronomers (Pence, 2009). FITS files contain both an image and a header that contains data about the image. The FITS structure influenced the design of VOTable, and the existing FITS keywords were incorporated into VOTable (Kent, 2002). The final result is that VOTable can either contain a FITS file or completely re-encode the data found in a FITS file (Ochsenbein et al., 2004, § Data Storage: Flexible and Efficient, ¶ 1). One of the intentions of VOTable was to make VOTables and FITS easily exchangeable so that data can convert from one format to the other without losing any information (Ochsenbein et al., 2004, § What can VOTable do but not FITs?, ¶ 2).

The Virtual Astronomy Multimedia Project (VAMP) has created another metadata standard, the Astronomy Visualization Metadata Standard (AVM). AVM is designed for education and public outreach; the focus is on print-and screen-ready “astronomical imagery, which has been rendered from telescopic observations (also known as “pretty pictures”) (Virtual Astronomy Multimedia Project). This content is conceived of as part of the NVO's Education Public Outreach (EPO), intended to reach the K - 12 community, "allowing user communities to build unique knowledge of the science behind the images" (NVO SDT, 2002, p.35).

Context of Use

The NVO provides a framework for accessing the global astronomical data repositories, which is complemented by an array of analytical tools. It allows remote access to massive data sets consisting of instrumental data from land- and space-based observatories, such as the Very Large Telescope array in Chile and Hubble Space Telescope, and from space missions, such as by NASA, as well as supports wide-spread access to project data, such as the Two Micron All Sky Survey (2MASS) and the Sloan Digital Sky Survey (SDSS). By providing access to individuals and institutions without access to telescopes and observatories, the NVO provides a "powerful engine for the democratization of science" (NVO SDT, 2002, p. 18).

One of the earliest adopters of networking and computing technologies (Krumenaker, 2001), the field of astronomy and astrophysics has embraced the VO concept and has worked diligently in the creation of tools and standards supporting the framework. Researchers are able to access the distributed data sets and use mosaicing services like MONTAGE to create enhanced images (by stacking and stitching multiple images) by utilizing the infrastructure and computing power of the TeraGrid.

Via the NVO website, users are able to perform a wide variety of functions; following is the list of services immediately accessible through the website (www.us-vo.org):

- Collect all data from a given position,
- Count matches between catalog entries and given positions,
- Query databases and cross-match object lists,
- Find data collections and catalogs by searching their descriptions,

- Integrate data from multiple positions and data sets,
- Query the VO from the command line,
- Convert text tables to VOTable format; analyze or visualize a VOTable,
- Browse and analyze SDSS, 2dF, and your own spectra,
- Explore the Multiwavelength Sky in the Vicinity of Transient Events that have recently been observed,
- Make mosaics from 2MASS, DPOSS, or SDSS images,
- Repair Image Coordinates in images with inaccurate or misaligned coordinate systems,
- Find, use, store, and edit sky footprints, and
- Perform Source Extraction and Object Identification by detecting objects in your own images and matching them with objects in the major survey catalogs.

Very early into the VO initiative, tools were already proving successful. In the abstract of "Discovery of Optically Faint Obscured Quasars with Virtual Observatory Tools", Padovany, et. al. (2004) states, "This work demonstrates that VO tools are mature enough to produce cutting-edge science results by exploiting astronomical data beyond "classical" identification limits with interoperable tools for statistical identification of sources using multiwavelength information."

Stated throughout NVO literature, it is intended to be used by astronomers and astrophysicists of all backgrounds, from expert to novice. At the time of this writing, the NVO is most accessible to the technically-savvy experts. A more user-focussed interface is planned for the next major initiative: the Virtual Astronomical Observatory. Long-term goals encourage access by students and the general public, promoting information technology and science literacy and creating "citizen scientists" (NVO, 2009c). An integral component of the Virtual

Observatory initiative is the establishment of an Education and Outreach Program (EPO). The EPO provides teaching resources for kindergarten and up, research fellowships and opportunities across pre-college and post-doctorate levels, and digital images to the arts and entertainment community (NVO SDT, 2002, pp. 35-36).

The 2006 NVO advisory report states,

> The NVO is already a valuable tool for researchers, and some research papers enabled by the NVO are beginning to appear. It would be useful to track these, perhaps by asking authors to notify the NVO of publication. The US journals are beginning to offer authors the option of indicating which facilities were used to complete the research, and NVO might ask whether it could be included as a facility to make the tracking easier. (p.4)

The NVO presently requests that scholars acknowledge the use of NVO tools and applications and provides suggested language for use on the NVO website (http://www.us-vo.org/).

It is possible to locate a range of scholarly research stemming from the NVO. The NVO conveniently cites publications and presentations in the annual reports, most of which reference NVO tools and capabilities. Over sixty were cited in the October 2007 - September 2008 Annual Report. Journals and conferences include Sky and Telescope, International Science Grid This Week, IEEE Transactions on Knowledge and Data Engineering (TKDE), the Astronomical Data Analysis Software and Systems (ADASS) annual conference, the Microsoft e-Science conference, American Astronomical Society (AAS) annual meeting, the American Association of Physics Teaches (AAPT) annual meeting, to name just a few.

Publications are also posted on the NVO website, as well as the European Virtual Observatory's website (www.euro-vo.org). Following are title examples: "Luminous AGB stars in nearby galaxies. A study using virtual observatory tools." "A Population of Compact Elliptical Galaxies Detected with the Virtual Observatory." "The Infrared Afterglow of Supermassive Black Hole Mergers." "Environments of Starburst Galaxies Diagnosed with the NVO." "Discovery of optically faint obscured quasars with Virtual Observatory tools."

While the field of Astronomy and Astrophysics is relatively small, those involved with the VO initiative recognized the need for widespread inclusion and participation across science, information and technology fields. It is worth nothing that the principal investigators of the original funding source hail from both the computing and astrophysics fields. The NVO Science Definition Team provides examples of involvement by surrounding fields. "Statistics and Computer Science disciplines have found astrophysical data to be of particular interest because of the size, complexity and its non-proprietary nature" (p. 9). The NVO faces similar technological challenges as to those found in the fields of high energy physics, computational genomics, global climate studies, geophysics, oceanography, etc. The methodologies and techniques incorporated in the development of the NVO will prove useful in many different scenarios, as other fields are dealing with similar issues of large data archives, distributed computational grids, and management of structured digital information (p. 18).

Access and Tool Development

The three primary components of the NVO are registry services, data services and computing services. The registry services are the most simple; they "allow services and other entities to be published and discovered" (NVO, 2004, p. 7). The data services involve the simple

delivery of data, be it a FITS file or VOtable. Computing services are the most complex and were not fully developed in the first few years of the NVO.

In order for any of the services to function or for the tools from the NVO to be useful, the data in the registries must be compliant with IVOA protocols. Those primary protocols include the Simple Cone Search Protocol (SCS), Simple Image Access Protocol (SIA) and Simple Spectral Access Protocol (SSA). Currently, each registry is identifiable by what protocol is used for the data associated with it. In 2006 work began on a new Table Access Protocol (TAP). The goal for TAP is that it will eventually replace the SkyNode and cone search services and become the primary VO interface for access to tabular data. (NVO, 2007, p.14)

Tools developed by the NVO

There are several access points to the data in the NVO registries, the NVO acknowledges that "no single interface or tool was likely to satisfy even a simple majority of astronomers" (NVO, 2007, p.1). The NVO portal uses XSL to interact with the XML results that are discovered via their search interfaces (Robert Hanisch, personal communication, November 23, 2009).

DataScope.

The most basic way to query the registries is DataScope. This service is a way to search for data for a single location throughout all of the NVO registries (NVO, 2004, p. 15).

Open SkyQuery.

A more complex way to search the registries than DataScope, this service relies on the registered data provided by catalogs of SkyNodes (NVO, 2004, p. 27). Using SkyQuery requires some understanding of SQL to query the SkyNodes but this approach provides a high level of control over search parameters for cross matching against all of the SkyNode catalogs. Also, there is an option to upload and cross match data in either VOTable or MS DataSet format (NVO, 2009b).

WESIX.

WESIX also uses the SkyQuery protocol (NVO, 2005, p. 31). This service does "source extraction and crossmatching for any astrometric FITS image" (NVO, n.d.). WESIX remotely runs SExtractor to execute source extraction. Also, "the resulting catalog can be cross-matched with any of several major surveys, and the results returned as VOTable" (NVO, n.d.). Once catalogs have been selected in WESIX, applets for Aladin or VOPlot can be run to visualize the data.

VIM.

According the the NVO, The Virtual Observatory Integration and Mining (VIM) service has access to most of the astronomical data in the world (NVO, 2008a, p. 28). This service allows astronomers to perform multiple cone searches at once and it is capable of creating image cutouts from any of the SIAP-compliant resources (NVO, 2007, p. 28).

VOEvent.

One of the newer services from the NVO allows astronomers to subscribe to astronomical events and track them over time. The protocol for this service is VOEvent, astronomers can download software to use this protocol and then track particular events as they are surveyed (NVO, 2006a, p. 35).

Inventory.

This interface finds all the datasets for catalogs and images for any region in the sky based on user query. Within the interface users can send the found data to the NVO Table Viewer or VIM for futher analysis. As of 2008, the data that is accessible via the Inventory includes "all Vizier, HEASARC, and IRSA catalogs; IRSA, Spitzer, and some MAST image sets; and will eventually cover all datasets registered with the VO (not all of these currently provide full data access services)." (NVO, 2008a, p. 27)

VO-CLI.

First released in 2006 (NVO, 2006b, p. 16), this is downloadable software for command-line access to search the NVO registries (NVO, 2007, p. 31). This software is the most extensible interface provided by the NVO, it can be used by application developers to access the data from the NVO, "Bindings of the VOClient API are available for most popular scripting and compiled languages and application environments, including Python, IRAF, C/C++, Java, FORTRAN, and so forth" (NVO, 2006b, p. 16).

Registry publishing tools.

The first software for publishing in NVO registries was the 2003 release of VORegistry-in-a-Box. This service was designed so that data providers could register their services with the NVO. Forms were included to describe data using OAI (NVO, 2003b, p. 9). As of 2005, VORegistry-in-a-Box was no longer being developed. Instead of using VORegistry-in-a-Box, publishers would need to register at the STScI registry page (NVO, 2005a, p. 13) or with Carnivore, the new registry publishing software.

Carnivore is frequently updated software for publishing, searching and harvesting services in the NVO (p. 39). This software uses Ajax and is kept up to date with IVOA standards.

Summer School software.

The NVO Summer Schools (NVOSS) were an exercise for the NVO to extend their development tools to the astronomical community. In order for students to interact with VO data, software packages were compiled. The most current NVOSS software is still available for download and it allows developers to interact with NVO data via Java, Python, C#, IDL, IRAF and Perl (NVO, 2008b).

In particular, the first Summer School was a way for the NVO to beta test the registries before releasing them for public use; students used Carnivore to access the registries and created a new cone search (NVO, 2004, p. 14).

Tools Developed Outside the NVO

The data that has been federated by the the NVO is highly extensible; it is possible to develop software that will read the XML-based metadata. The protocols developed by the NVO and IVOA can be used to identify VO-compliant software that is developed outside the NVO framework.

Mosaicing.

The JAVA-based mosaicing service Montage is technically not an NVO tool, it has no NVO oversight but it is a useful astronomical tool (Robert Hanisch, personal communication, November 15, 2009). In 2002 IPAC received funding from NASA for the development of Montage and the initial software engineering plan and general design for Montage began (NVO, 2002c, p. 14).

In 2004 Montage software was installed with the TeraGrid (NVO, 2004, p. 9) and the Pegasus workflow system was combined with ROME software to support Montage. Together, it is possible to use Montage to retrieve mosaics from the datasets of the NVO, these mosaics use "image stacking" processes to create images. Whether or not the requested mosaics require Pegasus to map out mosaic workflows depends on the complexity of the computations (NVO, 2005a, p. 31). In 2005 "tens of thousands of jobs" were succesfully intitiated, tracked and completed using ROME and Montage. (Hanisch, 2005b).

Development of Montage at IPAC has been stagnant for the last two years but it is still able to process mosaic requests (Robert Hanisch, personal communication, November 15, 2009).

Java parsers.

Most of the services designed by the NVO return table-formatted data, since this data is in XML it can be accessed, parsed and modified via Java parsers such as Simple API for XML (SAX) or Document Object Model (DOM) (NVO 2006 Summer School, 2006). Using these APIs it takes only "a line or two of user code to take advantage of a library that exists in the environment" (Robert Hanisch, personal communication, November 23, 2009). Both of these XML parsing options are also supported by IDL (Interface Description Language).

Saada.

Saada creates databases of astronomical data in both FITS and VOTable format data (Saada, 2009). In addition to supporting the VOTable format, Saada is ambitiously pursuing support of the TAP protocol before the protocol becomes fully operational in the NVO registries.

ATpy.

One other piece of software in the proliferation of astronomical tools, this software can "manipulate tables of astronomical data in a uniform way" (ATpy, 2009). The data table formats supported include FITS, IPAC, SQL and VOTable.

Strengths and Weaknesses

"We will know that the NVO is a success when it is used daily by thousands of astronomers, educators and members of the general public - and that it is taken for granted, just like the most successful web-based information services today." (NVO SDT, 2002, p. 12)

The National Virtual Observatory has been more successful than other online projects that attempt to collect large amounts of data and provide access to them online. Hsu et al (2007) claim that, "Clearly, the biggest challenge in fostering VCs [Virtual Communities] is the willingness to share knowledge with other members." (p. 1) Compared to other repositories, the NVO has more successfully encouraged sharing among members. There are many strengths of the NVO project that contribute to their relative success. There are also some weaknesses that would bear improvement in order to advance the quality of NVO services.

Project Strengths

According to the Atkins report, cyberinfrastructure projects should enable collaboration and knowledge discovery through grid-computing, networks, data sharing, and tool development (2003, p. 12). The NVO has developed a robust cyberinfrastructure as well as a functional interface; their success is due in large part to the dedication of the community.

Foundation.

The NVO began with clear goals specific to the subject are of astronomy. From the beginning, the NVO team was composed of dedicated astronomers, computer scientists, and librarians, meaning that the project had a greater chance of success. Two factors ensure funding: specific goals that aligned with NSF and NASA initiatives, and demonstrated committment by the project team and the community-at-large.

Interface.

An important feature of the NVO portal is the access to tools for data discovery and analysis. The NVO has also developed an interface that is compliant with HTTP standards. Their online interface is searchable in a variety of ways, the tools are available for use inside a browser, or to download as an application.

Reputation.

Many stakeholders in the NVO have established name recognition; the involvement of such prestigious institutions as John's Hopkins University and California Institute of Technology contributes to a wide dispersion of information sharing. Collaboration fosters participation from many sources. The Summer Schools gave NVO developers and specialists the opportunity to create advocates and teach future professionals how to use the tools. This knowledge has organically spread throughout the community. Due to a strong reputation, and word-of-mouth in the astronomy community, the NVO has little need for collection development outside of participation by stakeholders. According to an email from R. Hanisch to the authors, the NVO has received submissions from most current publicly funded US observatories. (Personal Communication, Nov. 17 2009)

Infrastructure.

Of all the NVO strengths, the infrastructure is especially robust. In the white paper by Williams and De Young,

> The key to providing the VAO environment described above is interoperability—an infrastructure that can federate these archives through common and standard ways of describing and accessing data, much as the HTML standard enabled the web itself in the 1990s. (p. 4, 2009)

The NVO is interoperable between several databases and surveys of astronomy data collected from telescopes, such as the 2MASS, and Sloan Digital Sky Surveys. The NVO also follows a common metadata plan that is standardized but extensible. This method of describing data makes the it more searchable and accessible. The NVO team and the astronomy community agreed on these standards early on, and the standards were robust enough to enable extensibility. Although the NVO is a national iniative, the standards implemented in their infrastructure are compliant with international protocols.

Weaknesses

The NVO is a continuing initiative; the foundation could improve by securing permanent funding and creating a more usable interface. Additionally, solicitation of more analytical reviews of the portal and tools could solidify an already reputation.

Foundation.

The original committment from the NSF was for five years, although the NVO was able to extend the funding for seven years, the NSF grant is not permanent. This possible inability to maintain funding in perpetuity is a common problem facing online collections. The current funding model lacks long-term stability, forcing the digital curation projects to plan for the short

term. Cyberinfrastructure projects must develop a model to address their long-term mission and necessary funding needs.

Interface.

The interface of the NVO would benefit from further development. Many links to sites and documents are broken and the overall site usability needs improvement. Though a usability study was completed in 2006, but site needs continuous assessment. The following year's annual report indicated the NVO interface should encourage casual use by allowing general data discovery (NVO, 2006a, p. 22). If the intended audience is to include non-specialists, then the tools are too exclusive; currently it is difficult to navigate or browse without a prepared location or query. There is also a lack of visual marketing for educational or general audiences.

Reputation.

The NVO currently enjoys a trusted reputation, but few analytical reviews of the tools and interface are available outside the internal documentation of the NVO. Most outside sources mention the NVO as a successful project, but an in-depth analysis of tools and interfaces by external reviewers would be beneficial.

Conclusion

A 2002 report notes that scientific discovery is punctuated by “bursts of creative growth, which follow [the] introduction of major new technologies” (NVO SDT, 2002). The astronomy community believes that they are once again at the brink of such a moment of major discovery made possible by Virtual Observatories.

The National Virtual Observatory is an excellent example of a successful large-scale, online data collection. The initiative, a cyberinfrastructure for astronomical data, has benefited from a focus on a discrete type of data, and a small, close-knit, and enthusiastic community of astronomers. This technically-savvy community also had the foresight to collaborate with information specialists to create strong metadata standards. Through the work of the NVO, the IVOA protocols have been developed to define a standard for how this astronomical data is represented and transferred. One result of the federation of data based on VO protocols has been the proliferation of tools that are able to read interoperable data.

The NVO shows flexibility to improve tools and services, and dedication to continue development through cooperation. In the future, the NVO could benefit from further metrics for success coupled with analysis and usability studies to hone the web interface, as well as a directed focus on long term data preservation.

Acknowledgments

A great deal of the information for this case study was drawn from guides provided on the NVO website, and from their own documentation. In addition, Robert Hanisch kindly and quickly answered many of our questions about NVO administration, structure, and other miscellany.

Appendix A - Milestones by year

1997 Astronomy and Astrophysics Survey Committee (AASC) is created

1999 "Astronomy and Astrophysics in the New Millennium" Panel Recommendations

2000 National Virtual Observatory Science Definition Team is created by the AASC

2000 National Virtual Observatory Interim Steering Committee is created

2001 National Virtual Observatory initiative is officially enacted with NSF funding

2002 International Virtual Observatory Alliance is formed

2004 First Summer School

2004 TeraGrid began full production

2005 Official release of the Resource Registry, DataScope and OpenSkyQuery services

2005 Second Summer School

2006 NVO "essential infrastructure" is complete

2006 Third Summer School

2007 Several standards, including Simple Cone Search, approved by the IVOA

2007 Montage deployed at IPAC

2007 NVO book based on Summer School materials published

2008 The Virtual Astronomical Observatory is created with joint NSF and NASA funding (the NVO initiative is folded into the VAO)

2009 White Paper: "The Role of the Virtual Observatory in the Next Decade"

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue ribbon advisory panel on cyberinfrastruture. http://www.nsf.gov/od/oci/reports/atkins.pdf

ATpy (2009). ATpy - Astronomical Tables in Python. Retrieved November 25, 2009, from http://atpy.sourceforge.net/

Babu, G. Jogesh, Djorgovski, S. George. (2004). Some Statistical and Computational Challenges, and Opportunities in Astronomy. *Statistical Science 19*(2). Retrieved December 1, 2009, from http://www.jstor.org/stable/4144415

Benson, K., Plante, R., Auden, E., Graham, M., Greene, G., Hill., M. (2009). IVOA Registry Interfaces 1.0. Retrieved from IVOA Documents October 21, 2009, http://www.ivoa.net/Documents/RegistryInterface/

Berriman, G., Curkendall, D., Good, J., Jacob, J., Katz,. D, Kong, M. (2002). Architecture for access to a compute-intensive image mosaic service in the NVO. *Proceedings of SPIE, 4846*(91). DOI:10.1117/12.461507

Brief Explanation of 2MASS. Retrieved from http://www.ipac.caltech.edu/2mass/overview/about2mass.html.

Choudhury, Sayeed. The Virtual Observatory Meets the Library. *Journal of Electronic Publishing 11*(1) Winter 2008. Retrieved December 1, 2009, from http://dx.doi.org/10.3998/3336451.0011.111

Deelman, Ewa et al. (2003). Grid-Based Galaxy Morphology Analysis for the National Virtual Observatory. SC '03: *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. Retrieved September 21, 2009, from http://portal.acm.org/citation.cfm?id=1048935.1050197&coll=GUIDE&dl=ACM&CFID=51171510&CFTOKEN=85903661

Graham, M. J., Fitzpatrick, M. J., & McGlynn, T.A. (Eds.). (2008). ASP Conf. Ser. 382, *The National Virtual Observatory: Tools and Techniques for Astronomical Research.* San Francisco: ASP.

Gray, Jim. The World-Wide Telescope *Science*, 293, 2001. Retrieved September 21, 2009 from http://www.us-vo.org/pubs/files/wwt.pdf

Hanisch, R. (2005). *Project Update*. Retrieved September 16, 2009 from the NVO Document Repository: http://www.us-vo.org/pubs/files/Project Plan Update Feb 2005.pdf

Hsu, M., Ju, T. L., Yen, C., & Chang, C. (2007). Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human Computer Studies, 65*(2), 153-169. Retrieved April 18, 2008, from http://portal.acm.org/citation.cfm?id=1222677.1223113.

International Virtual Observatory Alliance (n.d.). IVOA Registry of Registries. Retrieved October 29, 2009, from http://rofr.ivoa.net/

Kent, S. (2002). NVO metadata working group meeting: minutes of Aug 15, 2002. Retrieved September 16, 2009, from NVO Document Repository: http://www.us-vo.org/pubs/files/minutes2002-08-15.txt

Krumenaker, L. (2001, September). Virtual Astronomy. *Online*, *25*(5), 54. Retrieved September 25, 2009, from Professional Development Collection database. http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=tfh&AN=5119855&site=ehost-live

Lais, S. (2001, July 30). Virtual Observatory. *Computerworld*, *35*(31), 47. Retrieved September 15, 2009, from Computer Source database. http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=cph&AN=4974086&site=ehost-live

Mining the digital skies. (2000, June 3). *Economist*. Retrieved September 15, 2009, from Business Source Alumni Edition database. http://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bah&AN=3173823&site=ehost-live

Minutes of the Meeting of the Astronomy and Astrophysics Advisory Committee, 21-22 June 2004. Retrieved December 1, 2009, from www.nsf.gov/attachments/103128/public/minutes_2004jun.pdf

Moore, R. W. (2002). NVO architecture: Virtual observatory meeting. Retrieved October 8, 2009 from NVO Document Repository: http://www.us-vo.org/pubs/files/NVO-sys-garching.doc

National Research Council. (2001a). *Astronomy and Astrophysics in the New Millennium (Decadal Survey).* Washington, D.C. : National Academy Press. Retrieved December 1, 2009, from http://www.nap.edu/books/0309070317/html/

National Research Council. (2001b). *Astronomy and Astrophysics in the New Millennium: Panel Reports.* Washington, D.C. : National Academy Press. Retrieved December 1, 2009, from http://books.nap.edu/catalog.php?record_id=9840

National Science Foundation. (2001). *Award abstract #0122449: ITR/IM: Building the framework of the National Virtual Observatory*. Retrieved September 23, 2009 from: http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0122449

National Science Foundation. (2008a). *Cyberinfrastructure: A Special Report.* Retrieved November 29, 2009 from http://www.nsf.gov/news/special_reports/cyber/digitallibraries.jsp

National Science Foundation. (2001b). *Directorate for Mathematical And Physical Sciences Government Performance and Results Act (GPRA) Performance Report for FY 2001.* Retrieved from http://www.scribd.com/doc/1000152/National-Science-Foundation-GPRA-Rpt

National Science Foundation. (2008). *Program Solicitation NSF 08-537, Management and Operation of the Virtual Astronomical Observatory*. Retrieved October 1, 2009 from http://www.nsf.gov/pubs/2008/nsf08537/nsf08537.htm

National Virtual Observatory. (2002a). *NVO team meeting 30-31 July 2002 UIUC/NCSA Champaign-Urbana, IL*. Retrieved October 8, 2009 from NVO Document Repository: http://www.us-vo.org/pubs/files/2002jul30minutes.pdf

National Virtual Observatory. (2002b). *Building the framework for the National Virtual Observatory quarterly report April - June 2002*. Retrieved October 8, 2009 from NVO Document Repository: http://www.us-vo.org/pubs/files/fy2002q3.pdf

National Virtual Observatory. (2002c). *Building the framework for the National Virtual Observatory annual report October 2001 – September 2002*. Retrieved October 8, 2009 from the NVO Document Repository: http://www.us-vo.org/pubs/files/Y1-annual-report.pdf

National Virtual Observatory. (2002d). *Simple Image Access Prototype Specification: Version 1.0, September 30, 2002 (Updated 2002-10-06)*. Retrieved October 8, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/ACF8DE.pdf

National Virtual Observatory. (2003a). *Resource and service metadata for the virtual observatory version 0.7.* Retrieved September 16, 2009, from NVO Document Repository: http://www.us-vo.org/pubs/files/ResourceServiceMetadataV7.pdf

National Virtual Observatory. (2003b). *Building the framework for the national virtual observatory annual report October 2002 – September 2003*. Retrieved September 16, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/Y2-annual-report1.pdf

National Virtual Observatory. (2004). *Building the framework for the National Virtual Observatory annual report October 2003 - September 2004*. Retrieved October 17, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/fy2004AR.pdf

National Virtual Observatory. (2004). *Project Management Plan.* Retrieved November 29, 2009, from http://www.us-vo.org/pubs/files/fy2002q3.pdf

National Virtual Observatory. (2005a). *Building the framework for the National Virtual Observatory annual report October 2004 - September 2005.* Retrieved October 25, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY05AR.pdf

National Virtual Observatory. (2005b). *Science With the Virtual Observatory 2005 Summer School: Introduction to Java Libraries for use with the VO*. First accessed October 29, 2009 from: http://www.us-vo.org/summer-school/2005/presentations/web_libs/XMLparse.html

National Virtual Observatory. (2006a). *Building the framework for the National Virtual Observatory Annual Report October 2005 - September 2006*. Retrieved October 25, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY06AR.pdf

National Virtual Observatory. (2006b). *Building the framework for the National Virtual Observatory quarterly report April - June 2006*. Retrieved October 8, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY06Q3.pdf

National Virtual Observatory. (2007). *Building the framework for the National Virtual Observatory annual report October 2006 - March 2007*. Retrieved November 1, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY07ARWithSupps.pdf

National Virtual Observatory. (2008a). *Building the framework for the National Virtual Observatory annual report October 2007 - March 2008*. Retrieved November 1, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY08AR.pdf

National Virtual Observatory. (2008b). *NVO Summer School 2008 Software*. Retrieved November 9, 2009, from http://nvo-twiki.stsci.edu/twiki/pub/Main/SummerSchool2008/NVOSS2008-Software.html

National Virtual Observatory. (2008c). *How to Publish to the NVO.* Retrieved November 5, 2009 from http://www.us-vo.org/pubs/files/PublishHowTo.html

National Virtual Observatory. (2009a). *Building the framework for the National Virtual Observatory Quarterly report January - March 2009*. Retrieved September 16, 2009 from the NVO Document Repository: http://www.us-vo.org/pubs/files/FY09Q2.pdf

National Virtual Observatory. (2009b). *Open SkyQuery Help*. Retrieved November 22, 2009, from http://openskyquery.net/Sky/SkySite/help/help.aspx

National Virtual Observatory (2009c). *The Role of the National Virtual Observatory in the Next Decade.* Authors: Williams, Williams, and De Young, Dave. Retrieved November 22, 2009, from http://www.us-vo.org/pubs/files/Williams_VO_TEC.pdf

National Virtual Observatory. (n.d.). *NVO Core Application Flyers*. Retrieved November 22, 2009, from NVO Document Repository: http://www.us-vo.org/pubs/files/nvo_app_flyers1.pdf

National Virtual Observatory Interim Steering Committee. (2000). *Draft White Paper: Towards a National Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan*. Retrieved November 10, 2009, from NVO Document Repository: http://www.us-vo.org/pubs/index.cfm

National Virtual Observatory Science Definition Team. (2002). *Towards the National Virtual Observatory*. Retrieved November 5, 2009 from http://www.us-vo.org/pubs/files/sdt-final.pdf

National Virtual Observatory 2006 Summer School. (2006). *VO Client side Integration: Lessons Learned (C. Miller 09/11/06)*. Retrieved October 11, 2009 from http://www.us-vo.org/summer-school/2006/presentations/existing_environ_idl.html

NISO website. (2009) *Collections*. Retrieved November 11, 2009 from http://framework.niso.org/node/8

NVO Advisory Committee. (2002). *The report of the NVO Advisory Committee*. Retrieved September 16, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/nvoreport.pdf

NVO Metadata Working Group. (2002). Resource and Service Metadata for the Virtual Observatory. Retrieved Steptember 16, 2009, from the NVO Document Repository: http://www.us-vo.org/pubs/files/ResourceServiceMetadataV3.pdf

Ochsenbein, F. (n.d.). *VOTable documents.* Retrieved October 15, 2009, from http://cdsweb.u-strasbg.fr/doc-cds/VOTable/

Ochsenbein, F., Williams, R., Davenhall, C., Durand, D., Fernique, P., Giaretta, D. (2004). *VOTable Format Definition Version 1.1*. Retrieved October 14, 2009, from http://www.ivoa.net/Documents/REC/VOTable/VOTable-20040811.html

Padovani, P.; Allen, M. G.; Rosati, P.; Walton, N. A. (2004). Discovery of Optically Faint Obscured Quasars with Virtual Observatory Tools. *Astronomy and Astrophysics* 424, 545-559. Rerieved from November 29, 2009, from http://dx.doi.org/10.1051/0004-6361:20041153

Pence, W. D. (2009). *The FITS Support Office at NASA/GSFC*. Retrieved October 14, 2009, from http://fits.gsfc.nasa.gov/

Plante, Raymond L. *A Scalable Metadata Framework for the Virtual Observatory*. Garching confere, 2002. Retrieved September 21, 2009, from http://www.us-vo.org/pubs/index.cfm

Rots, AH. *Space-Time Coordinate Metadata for the Virtual Observatory (STC)* http://www.ivoa.net/Documents/latest/STC.html Version 1.33, IVOA Recommendation 30 October 2007

Saada (2009). *You Have Data Files - Saada Makes your Database*. Retrieved November 25, 2009, from http://saada.u-strasbg.fr/saada/

Szalay, A. S. The National Virtual Observatory. *Astronomical Data Analysis Software and Systems X, ASP Conference Proceedings*. 238. Retrieved September 21, 2009, from http://adsabs.harvard.edu/full/2001ASPC..238....3S

Tenopir, Carol, King, Donald W., Boyce, Peter, Grayson, Matt, & Paulson, Keri-Lynn. (2005). Relying on electronic journals: Reading patterns of astronomers. *Journal of the American Society for Information Science and Technology,* 56(8), 786-802. Retrieved December 1, 2009, from http://www3.interscience.wiley.com/cgi-bin/jhome/76501873

US Virtual Observatory Consortium. (2009). *The Role of the Virtual Observatory in the Next Decade*. Retrieved October 14, 2009 from NVO Document Repository: http://www.us-vo.org/pubs/files/Williams_VO_TEC.pdf

Virtual Astronomy Multimedia Project. *Astronomy Visualization Metadata Standard.* Retrieved November 29, 2009 from VAMP Web site: http://www.virtualastronomy.org/avm_metadata.php

Web and Grid Services Working Group. (2009). *IVOA Support Interfaces Version 1.0-20090825*. Retrieved October 10, 2009, from IVOA Document Repository: http://www.ivoa.net/Documents/latest/VOSI.html

Williams, R. (2002). *Grids and the virtual observatory*. Retrieved September 23, 2009, from NVO Document Repository: http://www.us-vo.org/pubs/files/vogrid.pdf

Williams, R., Ochsenbein, F., Davenhall, C., Durad, D., Fernique, P., Giaretta, D., et al. (2002). *VOTable: A Proposed XML Format for Astronomical Tables*. Retrieved September 16, 2009 from the NVO Document Repository: http://www.us-vo.org/pubs/files/VOTable-1-0.pdf