

# The Report of the NVO Advisory Committee

## December 11-12, 2003

### Executive Summary

The NVO Advisory Committee met for the second time on December 11-12, 2003, at Johns Hopkins University in Baltimore. Committee members in attendance included Gerry Gilmore, John Huchra, Sid Karin, Rob Kennicutt, Carl Lagoze, Paul Messina, Eve Ostriker, and Sidney Wolff. Presentations were made by key members of the NVO team.

The Advisory Committee was very positively impressed by the continued progress made by the NVO team. In particular, we highlight the following achievements:

- ✍ International agreement on many of the standards for describing and accessing data
- ✍ Progress on the definitions and practical implementation of registries, again in collaboration with the International Virtual Observatory Alliance
- ✍ Development of a system architecture with enough redundancy and flexibility that it should be able to adapt to the rapid evolution of technology, grid computing, and web services
- ✍ Effective and wide dissemination of information about the NVO to scientists and programmers concerned with the development of algorithms, software, and software systems for working with astronomical data
- ✍ The development of tools that have enabled demonstration science: rapid acquisition of multi-wavelength data about a small area of the sky to identify gamma-ray burst hosts; federation of data from surveys at different wavelengths to identify outliers in color-color space, a capability that resulted in the discovery of a new brown dwarf; and examination of galaxy cluster populations and galaxy morphology as a function of galaxy density, x-ray emission, etc. through combining multiple data sets and using grid technology to carry out the calculations

We make several recommendations for the future development of the project:

- ✍ NVO users should be engaged more actively in planning and deployment of the NVO and in advising on priorities for developing tools for interacting with NVO data; some services should be made easily accessible for use by the broader astronomical community.
- ✍ Other federal agencies, including especially NASA and DOE, should be kept well informed about the progress of the NVO and their relevant datasets should be accessible and compatible with NVO standards.

- ✍ With the basic NVO framework in place, consideration must be given explicitly to establishing standards for the included data: its provenance, quality, and integrity along with issues such as intellectual property rights and citation guidelines.
- ✍ Progress in the education and outreach activities has not kept pace with the technical progress. These activities should be re-examined in order to establish a long range plan with clear priorities and a greater emphasis on leveraging NVO resources in this area through partnerships with organizations with a broad reach and established track record.
- ✍ The NVO should develop an operational model, including budget, in order to estimate what it will take to create and operate the NVO as a long-lived “observatory” with a large user community.

## 1. Achievements

The NVO project has the goals of

- Providing a digital representation for the entire sky that can be used by astronomers, educators, and the public
- Providing standard web services for manipulating image archives and star catalogs
- Providing tools for processing entire collections.

The committee is pleased to see the substantial progress made by the NVO consortium toward realizing these goals during the year since our last meeting. The adoption of standards, the definition of registries, the selection and implementation of the first applications, the involvement with developing technologies, and the close and active collaboration with the International Virtual Observatory Alliance (IVOA) are areas where impressive progress has been made. The committee accepts the NVO annual report as a valid description of a successful year.

Standards are an essential prerequisite for community acceptance, implementation, and extension of the NVO products, and standards must be in place for the NVO to become a viable and important component of the infrastructure that supports the research community. It is an essential aspect of NVO that the datasets accessed through it be distributed internationally. Thus, it is particularly important that early international agreement on standards be achieved. Through the positive and collaborative approach of the NVO and IVOA teams, this agreement has already been developed to a high degree (VOTable being an excellent example). Additionally, the technical skills available in the NVO project have ensured a leading role for the NVO project in defining these standards inside IVOA.

The mission of the IVOA is to facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory. An example of the impact of IVOA is the adoption of IVOA identifiers by NASA’s Astrophysics Data Centers Executive Council (ADEC). The committee trusts this will be an example of NVO-NASA cooperation and that there will be successors.

The IVOA is not only proving to be useful in gaining agreement on international standards and building blocks for virtual observatories in astronomy, it is likely to inspire other scientific communities to follow suit in their fields. International standards organizations are quite difficult to create and operate, and the NVO and its founding partners ASTROGRID and AVO have succeeded. The member organizations now span the globe. See <http://www.ivoa.net/>.

While recognizing that a considerable commitment of time and resources by the most skilled people is required for the definition of standards, the committee endorses continuation of this effort. The ongoing work to define satisfactory standards for data models and especially for metadata is a major task, but well worth pursuing at high priority.

An application of standards that is critical for NVO development, and which is technically extremely challenging, is the definition, and practical implementation of registries. The committee was very impressed by the considerable progress made in this key development in close agreement with the IVOA. The plan to demonstrate an operational prototype at the January 2004 meeting of the AAS is clear evidence that the project has focussed its resources on key generic requirements and is able to deliver.

The committee explicitly supports the philosophy of adopting available tools and resources wherever appropriate (e.g., OAI), and focusing NVO resources on the provision at an early stage of practical tools/interfaces for data providers built on these available resources. This rapid implementation has implications for data integrity, as we discuss below, but underpins realistic access to international archives.

The committee is pleased to see that early demonstration science applies and extends key infrastructures that will enable wider future developments. Of specific note, we commend the Data Inventory Service (DIS), which clearly provides a sound basis for the future. Multi-wavelength searches are expected to be of increasing importance in astronomy and are manifestly a critical challenge for NVO. The choice of an interim application of immediate relevance to a field as topical as that of gamma-ray bursters, while at the same time developing technical capabilities relevant to many future implementations, is sensible.

The long term goals of the NVO listed above are excellent and ambitious. The technologies that are necessary to achieve these goals are far from mature and evolving rapidly. Even where there are *de facto* standards, some compete with others, and it is far from obvious which software services and interfaces will become the dominant ones. The NVO team has addressed these uncertainties extremely well. As recommended by the Advisory Committee last year, the project has conducted a risk assessment and developed a system architecture that should be able to accommodate the evolution of technologies. The NVO approach is innovative and promising. It uses technologies from the web services, the digital libraries, and the grid communities, with enough redundancy and flexibility that if a technology loses support, others will supply the needed functionality.

The NVO team also recognizes that they need to track and participate in the evolution of the technologies, and a number of NVO investigators are active participants and leaders in relevant research groups and working groups.

## **2. Community Engagement**

In its first two years, the NVO team has made significant strides in engaging the US and international astronomical community in the NVO project. These efforts have included development of the NVO public web site, contribution to documentation of standards at the IVOA web archive, technical training at an international astronomical software conference, and demonstration of application prototypes at the AAS and IAU meetings.

### *Software Scientists*

The outside engagement efforts of the team in its second year were principally directed at the community of astronomical software scientists. These scientists bear front-line responsibility for developing and maintaining the software tools that observational astronomers use for most of their research activities – obtaining, reducing, browsing, analysing, and archiving data. In many university departments and research facilities, the primary mechanism for disseminating new software tools to the community is via informal training initiated by the local software "gurus." Engaging this group of knowledgeable and technically proficient scientists is therefore crucial to the long-term success of the NVO project.

The first major public forum for introducing VO tools to the astronomical software development community was a VO Tutorial held at the Astronomical Data Analysis Software and Systems (ADASS) meeting held in Strasbourg, France in October, 2003. This annual conference is the primary forum for scientists and programmers concerned with algorithms, software, and software systems employed in the acquisition, reduction, analysis, and dissemination of astronomical data. The interest in, and impact of, VO within this community was evident in the attendance of more than 200 scientists at this tutorial session, at which technical aspects of three key VO tools were described and demonstrated.

In the view of the Advisory Committee, the NVO team's efforts to capture the interest of the software scientist community have been very successful. To capitalize on that constituency's interest and expertise and to leverage its considerable influence in the larger community, it will be important to maintain easily accessible, up-to-date web archives of tutorials, specification standards, prototype tools, and other VO documentation. In addition to recruiting events at ADASS, more intensive sessions for experienced VO service developers and resource curators would be valuable.

The NVO team has identified ramping up recruitment and training of graduate students and postdocs as a key goal of its third year community engagement efforts. The team is planning a summer school program staffed by NVO scientists, such that participants will be able to return to their home institutions with practical hands-on experience with NVO tools. In the view of the Advisory Committee, this and future efforts of its kind will be crucial in building a coalition of energetic users and

advocates for the NVO. As for the group of professional software scientists, provision of web-based documentation will be essential in enabling summer school “alumni” to pursue personal VO-related projects, as well as to provide “missionary” advocacy for NVO.

### *NVO Users*

The NVO has made remarkable progress in executing its plan to create a useful large-scale computational infrastructure for the astronomy community. The work to date has been in keeping with the originally proposed effort and has been guided nearly entirely by the project PI and co-PIs. It is now time to actively engage the user community in order to ensure acceptance and success. The user community should acquire a sense of ownership in the NVO and see itself as part of NVO, rather than viewing NVO as some sort of external service being provided by a third party. In particular, it is desirable to have community advice about priorities for developing useful services and tools.

As one area where community advice on priorities might be valuable, we cite the NVO theory-related efforts. It is the committee’s view that in the future it would be better to consider more general applications than the globular cluster demo we were shown. For example, since there is a very large community of theorists who perform hydrodynamic simulations (for models of cosmology, galaxies, star and planet formation, etc.) with either finite-volume methods or SPH methods, there might well be a great deal of interest among observers in a service that enables importation of a simulation data set, which could then be rotated, projected, redshifted, etc., in order to compare theoretical structures with multi-wavelength observations. Such a capability might have the added benefit of encouraging (and aiding) theorists to adopt standard data formats.

There are at least two non-exclusive approaches that have been used successfully elsewhere in order to engage the community. The first is to create and empower a board of users to guide the development and implementation of the NVO. This is a body that is considerably different from an external advisory or review committee. The idea is to create a board of insiders, i.e., a board that is structured as a functioning part of the NVO, and not a group of outside critics (even in the best sense of critic). In the pre-PACI days of the San Diego Supercomputer Center, this board was called the SDSC Steering Committee. Post-PACI, it was the NPACI Executive Committee. There were representatives from the approximately two dozen institutions that supported the original proposal to the NSF. The group was initially set up with working computational scientists as members, and with eternal vigilance it was possible to maintain that character of the membership. (It was necessary to avoid a tendency to slip toward delegating membership to user support level people. It proved necessary to establish a separate body to coordinate and deliberate issues related to user support.) The steering committee rarely initiated ideas; rather it deliberated and advised with regard to the ideas of the center’s management and usually was supportive. This support helped to guide the project’s decision making, and it also gave the decisions considerable weight at the funding agency. Furthermore, the group was effective in communicating the sense of the community on controversial issues relating to funding agency actions and to the project itself, far

more effectively than the self-interested PI could ever do. We recommend that the NVO consider setting up a similar group to function within the NVO context.

A second approach is to identify prominent individual members of the astronomy community and to approach them directly with offers to support their efforts with NVO resources, including labor. This is obviously not scalable to the entire community, but it has been proven in other contexts to be of great help in demonstrating the value of the resource to the larger scientific enterprise. This is the approach first taken by the Pittsburgh Supercomputer Center and was enormously successful. It is important that the selected users are well respected leaders in the field and that the fraction of total NVO resources so allocated be relatively small. Nevertheless it is reasonable to expect large impact from such a program. We recommend that the NVO consider identifying a small number of such individual PIs and initiating discussions leading to specific joint efforts to incorporate NVO capabilities into their research programs.

The Advisory Committee believes that an enhancement of the NVO's web presence could produce significant serendipitous yields in broader community awareness and engagement. This web enhancement could include reorganization of the NVO's own home page and seeking partnerships with widely visited web sites amenable to establishing NVO links. Especially among students and postdocs, web browsing is a frequent and low -financial-cost mechanism for learning about research developments outside of their home institutions. By reaching out to this large group, NVO may recruit many scientists at formative stages of their careers.

In a reorganization of the NVO web page, it would be useful to segregate categories of links by the expertise level of the intended audience segment. For example, a category for "Developers" could include much of the information linked under the current "Project" heading on the NVO page, and a category for "Users" could focus on applications. For maximum demonstration of the power and flexibility of the VO concept, the links on the NVO web page could comprehensively include both high-level NVO-team developed services, portals, etc. (e.g. Montage, Data Inventory Service, etc.), as well as data and analysis services developed by other IVOA partners (e.g. VOPlot, Aladin, etc.). Many future service developers will initially be drawn to the NVO as users of "shrink-wrapped" applications.

Given the relatively short timeline of the current NVO project and the need for generating a wide community base, the Advisory Committee recommends a web site upgrade as a high priority in the next year. This recommended upgrade is mainly functional rather than aesthetic: astronomers are accustomed to--and often prefer--working with "research-grade" rather than "commercial-grade" products. In addition, the Committee recommends deploying new beta applications on the web as soon as they become available. If a system of logging accesses to applications is implemented at this stage in the project, it will provide valuable feedback to the NVO team on which types of tools are most interesting to varying communities. Web access records will be useful to the team in deciding how to focus community-engagement efforts in subsequent years, and these logs will also provide concrete evidence of the project's impact for future agency reviews.

### **3. Engagement of Federal Agencies**

The astronomical system of access to data described in the 2001 Astronomy and Astrophysics Decadal Survey (McKee & Taylor, "Astronomy & Astrophysics in the New Millennium," NRC Press) essentially relies on a seamless interface between the agencies that support astronomy. The vast majority of archived US astronomical data and metadata (e.g. the ADS) is currently under NASA aegis. NASA is therefore a major stakeholder in the virtual observatory process.

We believe that it is extremely important for the NVO to keep NASA fully informed about, and engaged in, both the continued maintenance and development of the archives and the development of the NVO itself. This is also true for the DOE, which has played a major role in at least one ground-based astronomical survey, the Sloan Digital Sky Survey, and may play a major role in the Large Synoptic Survey Telescope.

There has been a concerted government and community effort in the past few years to foster interagency cooperation and collaboration in astronomy. Pushed by both the OMB and Congress, there is now a FACA level interagency committee, the Astronomy and Astrophysics Advisory Committee (AAAC), charged with fostering interagency collaboration to improve research efficiency and cross fertilization. For more than a decade, there has also been a community based NAS/NRC committee, the Committee on Astronomy and Astrophysics, whose role is to shepherd the recommendations of the astronomy and astrophysics decadal surveys. Given the inherent interagency aspects of the NVO, we believe that the NVO should highlight its potential and its achievements in these and similar forums. The NVO should serve as a model for interagency cooperation in the field of astronomy and astrophysics and related physics and also in education and public outreach.

The NVO leaders should also consider organizing a briefing to the Director and Deputy Director of CISE to show CISE management the exciting NVO accomplishments and motivate CISE to pursue the creation of a powerful cyber infrastructure for the nation's research community.

### **4. Data Issues: Integrity, Curation, Bonding, Authorship**

The committee read with great interest the white paper "A Virtual Observatory Based on Publishing and Virtual Data," and discussed the accompanying presentation by Roy Williams. We are gratified to see that the project is confronting some of the issues associated with long-term archiving and access of the VO data, and how this information will be integrated with the astronomical literature.

Many aspects of a traditional publishing model are relevant to "publication" of datasets and virtual data, and the white paper identifies some of the most critical requirements for effective long-term publishing. These include the importance of "bonded" storage, that is, storage of the data in a library that commits to providing long-term access, its attendant inquiry services, and migration of data formats as standards evolve. This principle of bonding has already been embraced by the major refereed journals in astronomy, as the scientific societies that own them have

committed to guaranteeing long-term access and format migration as needed. As pointed out in the white paper, many options exist for bonded storage of VO data, including large data centers, journals, university libraries, or warehouse sites set up explicitly for this purpose. Our committee firmly endorses the establishment of bonding as a requirement for including datasets in the NVO. Since this implies a commitment on the data curators that will extend well beyond the life of this NSF grant, it is important that the bonding requirements and principles be firmly established and agreed to by the VO partners in the near future.

An item that the committee would like to see given more explicit consideration is that of the integrity of data sources accessed by the NVO services. An admittedly delicate issue, but a critical one, is the level of requirements placed on data quality, integrity, and curation for inclusion in the VO. The white paper describes several types of publishers, ranging large established data centers and major projects to individual scientists. The VO as currently planned would almost certainly incorporate some combination of these components. From the presentations at the review it seems likely that strict standards of data formatting (and perhaps curation) will be enforced, but it was unclear whether similar requirements on accuracy and integrity of the data will also be imposed. Indeed we sensed that there is considerable range of opinion on the latter within the VO consortium itself. This is a serious concern, because in the absence of consensus and clearly articulated standards, the natural migration will tend toward an open and uncontrolled system, with potentially debilitating effects on the use and integrity of the VO data collection as a whole.

The philosophy adopted at present delegates responsibility for the integrity and reliable description of a data archive to the curator of that archive. While noting that this is one possible approach, the committee also notes that each NVO user is responsible for the scientific interpretation of any datasets accessed by NVO. For any user to be confident in any scientific conclusion based on a complex query, that user must be able to know exactly which data archives were accessed and be able easily to find any relevant quality-control information describing every data set. It is probably essential that a user be able to select/deselect specific datasets from all those identified by the NVO processes as being potentially analysed for a query. The committee believes that information to allow such a decision must be collected and provided as an integral part of the NVO query interface.

We urge the project to confront these difficult issues and develop data quality standards and a plan for enforcing them over the coming year.

During the presentation a number of other prerequisites for effective long-term scientific use of the data were described. These include "provenance" (the capability to trace the data to their original source), and the ability to document and preserve (or regenerate) multiple versions of datasets. The concept of virtual datasets may be of great value here, because they allow for the storage of operations performed on a dataset (such as recalibrations and other modifications to the data), without the need to corrupt the original data source. This latter area is especially important for the astronomy literature, because the ability to reproduce published results is a fundamental tenet of the scientific process. As the sources of data flowing into a single project or paper become more complex in the VO era, the ability to trace back

to the sources of these data in an unambiguous manner will become even more important.

Another area that may need more attention concerns intellectual property and citation guidelines for real and virtual datasets. In traditional scientific publishing there are well-defined standards and procedures in place governing the peer review and publication processes. While these procedures are being adapted to ensure efficient dissemination of new results within the astronomical community, they operate within a within a pre-existing system of copyright law and professional rules for establishing scientific priority and intellectual property rights. These protections are implicitly incorporated when an article is published in a peer-reviewed journal, so we often take them for granted. However, when conceiving a new system for "publishing" data it is important to recognize that these implicit protections will not be in place, and it may be prudent to establish a rudimentary set of guidelines to ensure that proper attribution and credit are assigned to the publishers of such materials. To cite a few specific issues: How will someone cite a "published" VO dataset in the literature? Is it acceptable for an individual to reproduce large portions of a posted dataset (or many datasets) and republish it as their own? Will individuals be allowed to harvest all of the VO data in a given subject area and repost it as their own, or republish it in a copyrighted journal or book? If serious errors are identified in a large dataset that has been harvested from other primary sources, who bears the responsibility for correcting the posted data? If the original curator is unable or unwilling to do so, can a third party republish the corrected dataset as his or her own?

These latter concerns are not urgent, nor are they likely to fundamentally impede the ambitious vision that is articulated in the white paper. However recent experience with astronomy's one major foray into open-access self-publishing--the ArXive e-print server--raises some cause for concern. Although electronic preprints have not generally been regarded as original publications by our professional community, a growing number of scientists are demanding that their self-published preprints be cited in the literature on the same basis as peer-reviewed articles, and a few individuals have gone so far as to raise plagiarism accusations against colleagues who failed to cite and credit their self-published work. In the absence of any professional guidelines, we risk having the agenda of the debate defined by the most extreme fringes of our community. Hence it might be prudent for the VO leadership to consider these issues and establish a reasonable and fair set of citation and intellectual property guidelines from the beginning.

## **5. Education and Public Outreach**

It is the judgment of the Advisory Committee that the outreach and education activities of the NVO do not meet the high standards set by the other aspects of the project. As we stated last year, we believe that the task of the NVO is to create a framework that can support a variety of educational programs and outreach applications and not to carry out directly all of the functions enabled by the NVO. It is therefore crucial that leadership of the education and outreach programs be provided by someone with a broad understanding of the needs of the potential user communities, including both formal and informal education. The head of the NVO/EPO program should be committed to, and focus on, forming partnerships with groups already actively engaged in delivering products that have been shown to be

effective. The emphasis should not be on developing new delivery programs within the NVO itself. It was unclear from the EPO presentation what progress had been made during the past year, how priorities were set for EPO activities, and what partnerships were actually in place.

The committee agrees that a diversion of limited project resources and investigator attention to these areas is not the best way of solving these issues. A better choice would be to engage in a synergistic relationship with a project that has more expertise in outreach and education. The NSF-funded National Science Digital Library (NSDL) project is one example of a likely candidate for this collaboration. Technically, NSDL is intended to federate and provide access to a variety of diverse resources via common infrastructure. Organizationally, NSDL is intended to provide a center for technical innovation in a focus area (science education via the Web) and act as a community center for affiliated organizations. Collaboration between the NVO and the NSDL would seem to offer the opportunity to achieve a stronger NSDL from the technical perspective and a stronger NVO from the education and outreach perspective.

There are many other groups that might be engaged in a mutually beneficial collaboration. Another example at the 9-12 level is the Virtual High School, which has been partially funded by the NSF and which provides an accredited curriculum that results in high school diplomas. NVO could provide some of the “lab” facilities as well as some course content. The web site is <http://www.govhs.org/website.nsf>.

The committee recommends that the NVO reassess its education and public outreach program, re-evaluate the scope of what can be accomplished, and evaluate a variety of partnership opportunities. The end result should be a clearer master plan with a realistic assessment of what can be accomplished.

The committee recognizes that cross project collaborations are not always easy to effect. Many potential collaborations will be complicated by the fact that the funding is derived from different sources (for example, NSDL’s funding comes from EHR) with different program mandates. In those cases where the funding comes from different directorates within NSF, the respective NSF program directors should work to overcome bureaucratic impediments.

## **6. Future Planning**

The committee believes that the NVO has been successful in demonstrating overall feasibility but that it now should begin the progression to its next stage of existence, namely real operation for the astronomical community and the public. The NVO has put in place systems that are available or will soon be available to the astronomy community:

- ✍ Portals and web service interfaces to analysis procedures hosted by sites in the US and Europe
- ✍ Process management systems, i.e. data processing pipelines to create derived data products such as mosaics

- ✍ Web services that provide uniform capabilities across NVO catalogs and image archives, e.g., cone search, VOTable catalog query, and simple image access
- ✍ Data access layer: management of methods of data encoding formats for access based on physical quantities; some of the formats themselves are products of NVO (VOTable transfer format)

NVO plans to offer simple production services in year 3, which we endorse, but we recommended that NVO also develop a plan for obtaining the necessary resources (data archives, computing, network bandwidth) to support the production applications that will be released to the community in 2004. It would be tragic if the NVO services become so popular that they overwhelm the resources available and thus alienate the emerging NVO user community.

Looking further into the future, we believe the NVO needs to develop an operational model, including budget, in order to estimate what it will take to create and operate the NVO as a long lived “observatory.” Based on the experience of observatories that acquire data, it is clear that some core staffing will be required to maintain standards and documentation, deal with the issues of data provenance and bonding described above, keep the community informed about new acquisitions and capabilities, and provide at least limited user support.

One way to estimate the costs of this effort would be to write a hypothetical RFP for a service contract that would carry out the necessary tasks. The management expertise available at the NVO partner institutions could be used to develop the requirements for sustained operation. This information should then be provided both to the potential funding agencies and to the advisory apparatus so that resources can be identified to ensure that the NVO continues after the termination of the current grant. Given the long lead-time often required to put federal funding in place, this planning effort should be initiated during the next year.

In summary, the information technology and cyber infrastructure components of the NVO are making excellent progress, and there is a sound approach and plans for the evolution of the systems and services being put in place during the lifetime of the current funding. The very success of this effort, however, means that new demands will be placed on the NVO by a growing user community, and planning to accommodate these demands should be high priority both for the NVO collaboration and their sponsors.