

# The Report of the NVO Advisory Committee

## January 13-14, 2005

### Executive Summary

The NVO Advisory Committee met for the third time on January 13-14, 2005, at the San Diego Supercomputer Center. Committee members in attendance were Roger Blandford, Gerry Gilmore, Martha Haynes, John Huchra, Rob Kennicutt, Carl Lagoze, and Sidney Wolff. Presentations were made by key members of the NVO team.

The Advisory Committee was again very positively impressed by the continued progress made by the NVO team and by its response to the recommendations made in our previous report. An update on those recommendations follows:

- *Recommendation:* 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.
  - The NVO has established a Science Steering Committee; developed a number of useful applications; operated a summer school that trained astronomers in how to use the NVO and write simple programs to access compliant databases; redesigned the web page so that it is more oriented to users; and presented applications and some early science results at the AAS meeting in January.
- *Recommendation:* 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.
  - Both NSF and NASA are being kept well informed through reports, and both were represented at the meeting of the Advisory Committee. More effort will be required to engage DOE.
- *Recommendation:* 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.
  - Some progress has been made on defining quality metrics and making it possible to include data quality information in the metadata standards. However, this area, along with long term curation and availability of data, remains a challenge and is discussed in more detail in our report.
- *Recommendation:* 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.

- Several focused partnerships with institutions experienced in pre-college education and public outreach have been established, and the development of programs and educational modules is in progress. It is too early to assess the success of these initiatives, but the direction of the program is now clear, and we look forward to a report next year.
- *Recommendation:* 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.
  - A white paper (From Framework to Facility) on the transition from the current development phase to an operational phase was developed and submitted to both NASA and NSF. We were very pleased to hear that planning is in progress for a jointly funded continuation of the NVO after the current funding expires. Our report discusses some of the issues that both NVO and the funding agencies must address in order to ensure a smooth transition.

The primary new challenges we see going forward are:

- Managing the transition to the new phase of operations in such a way that the impressive success of the NVO to date and its corporate legacy are maintained
- Maintaining an appropriate balance between support of a growing user community and the continued development that is required in a rapidly changing technical environment
- Ensuring the integrity of the science results published on the basis of data obtained through the NVO, which in turn involves issues of data quality and maintenance of data files over long periods of time.

### **The Transition “From Framework to Facility”**

The primary challenge for the NVO is to make the transition to the next stage of the project and to the next stage of funding. The NVO is more advanced than the UK AstroGrid project in its thinking about the transition from “framework to facility.” In the paper on the issues involved in making such a transition, the NVO has identified key elements for a sustainable VO and made some initial estimates of costs. They itemize the following services that should be supported: User Support, Software Maintenance, Software and Systems Development, Data Provider/Publisher/Curation, Theoretical Simulations, Coordination with Computational Grids, Authentication and Security, Long-term Data Preservation, Links to the Astronomy User Community, Research Program, and an EPO program. These all would seem to be key elements in deploying and sustaining a usable facility that delivers a valuable research/access capability to the community. In the UK, certain elements of this agenda are being addressed in a generic way by the e-Science Core Programme--software maintenance, data curation and preservation, authentication, and security. These would seem to be common problems for

many research communities and NSF/NASA may wish to consider the potential benefits of some coordination and sharing of best practice among different communities. However, the US NVO seems well advanced in their thinking on these issues and is to be commended on their plans.

As we understand it, the next phase of the NVO will be funded jointly by NSF and NASA. Both agencies require that the funding for the next phase be granted on the basis of a fair and open competition that involves the solicitation of proposals through an Announcement of Opportunity; the proposals will be peer reviewed. We understand the responsive proposals will have to include strategic goals, metrics, priorities, and timelines for accomplishing the tasks specified.

While we recognize the importance and necessity of an open competition for the management of the NVO facility as it matures beyond the current grant, considerable danger to the project exists if the process becomes distracting to the project team, if there is a loss in continuity of the knowledge and capabilities already developed, or if the competition causes anxiety over employment issues. Therefore, this process imposes obligations on both the funding agencies and on the current management of NVO to ensure a smooth transition. We urge the agencies to conduct the solicitation and review as expeditiously as possible and to make sure that the impact on NVO staff productivity and morale is minimized. In particular, continuity of critical personnel and of their terms of employment needs to be maintained.

A general comment: one of the clear strengths of NVO is that it is virtual. This provides a diversity of knowledge, experience, and approaches. While the transition into a stable long-term state requires a structure, careers, employers, etc., and there are advantages in having a local critical mass for some aspects of projects, retaining the strength of the virtual is highly desirable. The management cost is a price worth paying for quality. Keeping the NVO participants distributed, with subcontracted work packages to the appropriate center (data, supercomputing, astronomical, computer science, industry...) is a model worth considering seriously.

Over the next few months, the NVO should work with NSF to define the corporate legacy and to develop a mutually acceptable set of deliverables and documentation that will be provided at the end of the first 5 years of funding in order to guarantee that this legacy is maintained through the transition to the "facility" stage of the NVO. These deliverables should serve to pass on information acquired during the current phase of the project to a potentially new operator or even to those responsible for operations under the current management, since it is likely that in either case at least some new personnel with a different skill set will be required.

In addition to a set of detailed documentation, we also recommend that the NVO prepare a white paper that summarizes both lessons learned and problems for the future, especially in the transition to an operational infrastructure. The committee believes that the experiences of the NVO project team in the creation, management, and sustainability

of the infrastructure they have created will be of great interest to other communities. The NVO is but one example of similar efforts in other scientific disciplines.

In general, the committee agrees with the project team that in its next phase the NVO will face a number of issues that are distinct from those addressed during the current research and prototyping phase. We believe, however, that it is essential that the characterization of the next phase as “operational” not exclude a requirement for a substantial amount of new development of technology and functionality in the NVO. It is an understatement to say that the NVO exists in a dynamic context of changing web standards, computing and network capabilities, and advancing scientific context. For the NVO to remain relevant, it must dedicate a substantial amount of future resources to moving its state-of-the-art forward.

### **Data and Metadata Quality**

The value of the National Virtual Observatory (NVO) is limited by the quality of the data it hosts. This truism presents a challenge to the management of the NVO. To what extent will the NVO attempt to police and provide an imprimatur for the various data sets that it will make available? This is not an issue for routine observations from major space observatories that have passed through rigorous protocols imposed by large user groups. Neither is it a question for rare instances of fabrication, incompetence, and plagiarism. However, there will be a large quantity of observations submitted from unfamiliar sources and inexperienced observers where the accuracy and integrity will be open to question.

Anecdotal evidence that this is already an issue is a comment by an early user of NVO. She had an ambitious project she wanted to do but discovered that first she had to invest substantial time in understanding each of the databases she wanted to use and what information was contained in the associated metadata.

A similar quality problem arises with the preprint servers where the solution has been to recognize that refereeing articles is impractical so that essentially all submissions are accepted. The Advisory Committee believes that a similar judgment is valid for NVO-accessed data. It is neither practical nor advisable for the NVO itself to attempt to validate most data submissions. We do believe that it is reasonable for the NVO, in addition to performing simple checks on compliance with NVO protocols, to require the publisher of data to provide enough information in the header files for the critical user to perform his or her own assessment as to the quality of the data, given its provenance. A flag that indicates the level of external review, in particular acceptance by a refereed journal, will also be valuable. It will be necessary to obtain the cooperation of journal editors to automate this.

In general terms, therefore, the committee continues to endorse the open source principle for data submitted to the NVO, but perhaps with the modifications suggested above and in the science commons project (<http://science.creativecommons.org/>), which describes a method to permit providers of data to specify allowed uses and to provide

details for proper attribution. The committee looks forward to a discussion of these matters at its next meeting.

The NVO architecture as described to the committee heavily relies on metadata created by, and collected from, distributed sources. As a consequence, the ultimate quality of the NVO depends on the accuracy and completeness of such metadata. Already, the project team has faced the issue of metadata quality and the need for improving it. This problem is not unique to the NVO and is a troubling reality for most OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) based implementations. There are no well-known easy solutions to this problem, outside expensive human intervention. The committee doesn't propose any specific solution, nor does it suggest that the project team define a solution. But, the committee does suggest that the project team factor these complications into its planning for an operational phase and understand the costs of maintaining metadata quality and the problems with not doing so.

### **The NVO and the IVOA**

The US NVO program is being conducted in what is necessarily both a national and an international context. This complex context requires careful consideration to ensure not only that NVO remains viable as a US program, but also that the full potential of IVOA is leveraged to supplement and complement national capabilities, strengths, and requirements. There will inevitably be compromises required among the various partners in IVOA, due for example to different regional funding priorities and different local community perceptions of the role of the Virtual Observatory. A healthy NVO in this context requires that this inevitability be designed into the structure currently being developed.

An obvious and dramatically successful example of this, well implemented to date, is the effort to develop international standards at the very many levels where they are required. There are other important aspects of NVO policy and NVO implementation that will need to be agreed in the same context. Some of these may have resource implications for the NVO, some may reduce NVO resource requirements, and others may simply require that NVO priorities cannot always be implemented as wished locally. Local reporting and review requirements will differ, as will local community expectations.

The issue of data quality and integrity may well require some more obvious compromises. A scientist using NVO will want to be certain that the databases he/she is using are stable and in this sense under the control of the user. Historically, databases have had some form of published, refereed description of their content. Implementation of that requirement in such a way that a responsible scientist can always repeat an experiment, with known information being provided with sufficient backwards compatibility to support a later refereeing process, remains to be fully integrated into the IVOA—and indeed there is considerable debate among the VO developers and potential VO users about whether and how to deal with this issue. The method for incorporating “quality flags” and preserving databases will have to be agreed internationally, so that the outcome of a specific VO “experiment” does not depend on the location of the scientist.

Agreeing a policy on what datasets are to be accessed and what user-override/selection process is put in place requires careful consideration. At face value, there may be a conflict with the goal of simplicity and ease of use of the interface, yet an agreed IVOA policy is needed. The alternative is GIGO (garbage-in, garbage-out).

One example of the way in which NVO has already benefited from IVOA involvement is in the distribution of tasks. Over the past year, some IVOA partners have concentrated more on infrastructure, standards, etc., while the US program made a significant effort to provide applications to the astronomical community. It would be worthwhile to consider negotiating with IVOA partners to attempt some further optimization of the global and national activities. This is in part already underway. With suitable effort, such coordination should significantly reduce the risk that NVO, and IVOA, could suffer from a funding challenge at one or a few participants and would make the whole process more robust.

### **Data Preservation**

The committee expects that the NVO will face increasing pressure to play a role in preservation of the data that it makes available through its service infrastructure. This will be especially true when the data and the results of NVO services that manipulate it are included in established publication venues such as journal papers. A fundamental principle of scientific research is that it should be possible to replicate experiments and results.

The issue of data preservation is complex and is, in fact, the subject of a number of research efforts. Therefore, while we do not expect the NVO to directly address this issue, we believe that there is an opportunity for the NVO to establish partnerships with university libraries, which are increasingly interested in digital curation of text and data. The traditional role of university libraries in preservation makes them ideal partners for this task. We suggest that such a partnership might be a model for future similar efforts within the realm of cyber infrastructure. To the extent possible, it may also be appropriate to transfer data sets to existing data archives, such as those maintained by NASA and the fledgling ground-based archive being established at NOAO. We also note that SDSC has for some time now expressed a very strong interest in providing data holding and curation services.

This problem may also be addressed at least in part at the international level. There are worldwide several major data centers that hold the vast majority of astronomical datasets. Many are specific—e.g., observatories, space agencies, etc., while others are more general. It may be possible as part of NVO/IVOA to negotiate for smaller data sets of high quality to be located at some of these dedicated centers irrespective of national boundaries. For example, it may be possible to deposit some data sets at centers such as CDS, CADC, etc. This would ensure stable curation and stable access for a useful timescale for IVOA. Active participation in the growing Data Centre Alliance activities will ensure this option remains viable and on the relevant agendas.

Over the course of this meeting of the Advisory Committee, an issue arose related to the need for the NVO to provide a document repository function. This is in addition to the data-oriented and service-oriented functions that the project team has already investigated. The committee recommends that this function could be provided through any number of existing institutional repository software packages including DSpace (<http://www.dspace.org>), ePrints (<http://www.eprints.org>), or Fedora (<http://www.fedora.info>). Regardless of the software chosen, the committee recommends that the project team include in their long-term plans the facilities for such institutional repository capabilities, which could then play a role in existing and developing institutional repository federations.

### **Authentication/Authorization issues**

The NVO has so far been concentrating on developing a set of services for the community without implementing any restrictions on access to facilities for different sets of users. Nonetheless, given the expressed intent to develop a “third leg” to the NVO that combines access to data and computing resources for data mining and analysis, the NVO staff recognize that they need to develop an adequate security process for the authentication and authorization of potential users. This will be a challenging issue to address because requirements of this kind are somewhat contrary to the very open access philosophy that has pervaded the NVO project so far. Perhaps as a consequence, and as the presentations made clear, the plans for authentication/authorization are not at a very advanced stage. In part, this is also undoubtedly due to the incomplete story on “AAA” from both the Web Services community and the Internet2 Shibboleth project. The NVO should investigate whether there are generic solutions developed elsewhere that meet the needs of the scientific community. There is also a need for dialogue between the NVO and the Grid and Internet2 communities. In addition, NSF and NASA may wish to contemplate the desirability or otherwise of a multi-discipline Certificate Authority. By the time of the next NVO review, the project should be in a position to present a more detailed plan for authentication and authorization of NVO users.

### **EPO**

The committee was pleased to see that major improvements in the NVO Education and Public Outreach (EPO) program have been implemented over the past year. Under the capable leadership of Carol Christian, the scope of the program has been reduced to a reasonable scale, and the project is exploiting partnerships with existing EPO efforts at Adler and elsewhere, both actions recommended by our committee in last year's report. Pilot projects now in progress cover a range of outreach activities, including developing metadata and a simple image access protocol suitable for use by the media; registering NASA EPO collections with the NVO along with querying services; and working with teachers to develop web-based curriculum modules for K-14 and college courses.

Our remaining concerns are relatively minor but are worth noting. While the project contains sufficient funds to support Christian's work in the program at the 50 percent level, her prior commitments have limited her involvement to 10 percent of her time.

While she has made remarkable progress in the last year, it was unclear to us whether this level of effort was sufficient to sustain the momentum she has developed. It was also unclear whether sufficient administrative support was being provided for the more routine accounting, assessment, and reporting functions associated with the program. We encourage NVO management to ensure that a sufficient level of administrative support is available, so that Dr. Christian can focus her limited time on the top-level organization and management of the EPO program.

### **Community Engagement**

Following the recommendation of last year's Advisory Committee report, the NVO team has taken several important steps to engage the community. The redesigned web page is more user friendly and informative. The sessions devoted to "Research with the NVO" at the 2005 January AAS meeting were excellent and well-attended. Notably, several of the oral presentations showed example science applications that demonstrated the power and accessibility of the NVO. The Summer School likewise appears to have been excellent and enthusiastically taught as well as received. All of these efforts lay the groundwork for educating the astronomical community in the potential and capabilities of the NVO. We have the following particular suggestions:

- a) The web page is very important. It needs to be maintained and updated as needed. It would be useful to keep track of its use.
- b) It is useful to document and advertise successful science applications of the NVO. It might be appropriate to try to organize similar sessions on "Research with the NVO" at some future AAS meetings, perhaps at the January 2006 meeting or during a more extended special session at the 2006 summer meeting. Such a special session could highlight both the tools available and some of the research already accomplished.
- c) We concur with plans for a second summer school and agree that the focus audience should be younger people (grad students and postdocs), regardless of their current expertise in relevant programming languages, environments, etc. Specifically, studies show that women tend to take the specific qualifications listed in advertised positions or opportunities literally, and may not apply if they seem not to meet the requirements for programming experience. Since the first summer school demonstrated that one could accomplish some significant tasks even without previous Java experience, we suggest that the announcement of the summer school be made as general as possible to encourage participation by a diverse group of astronomers. We note that the format for the school may continue to evolve in response to both experience and demand. In order to ease the workload on the NVO staff, we suggest that it might be useful to invite some of the participants in the 2004 school to attend the 2005 school as instructors/project coaches.
- d) We are very pleased with the progress made in 2004 towards engaging the astronomical community; at the same time, we urge the team not to devote too much effort to this activity, but rather to continue to give high priority to development.