

The Report of the NVO Advisory Committee

February 16-17, 2006

Executive Summary

The NVO Advisory Committee met for the fourth time on February 16-17, 2006, at Johns Hopkins. Committee members in attendance were Martha Haynes, John Huchra, Sid Karin, Rob Kennicutt, Carl Lagoze, and Sidney Wolff.

The Advisory Committee has followed the development of the NVO closely since its inception. During that time, the NVO team has succeeded in bringing together diverse expertise and interests in order to build a cohesive project. It has fostered a successful international partnership, succeeded in balancing technological innovation with real-world capabilities, and managed to keep NVO development science-driven. The NVO is no longer merely a *virtual observatory*. It has become a *vital tool* for the astronomical community.

A critical challenge for both the NVO team and the NSF is to ensure a smooth transition to the new agreement for managing the NVO. It is important that the momentum and expertise that have been developed to date not be compromised during the transition. We recommend that the NVO team give high priority during this final year of phase one funding to documenting the lessons learned, achievements to date, and vision for the future.

The Advisory Committee was again very positively impressed by the continued progress made by the NVO team during the past year. It appears that a good balance was achieved between providing tools useful to research astronomers and continuing work on the NVO architecture and infrastructure.

The summer schools have been very successful in training a cadre of people at various institutions in how to develop software for, and make use of, the NVO. These people should be encouraged to pass their knowledge on at their home institutions, including to REU students.

The strategy of forming partnerships with experienced education and outreach groups is appropriate but is dependent for success on the ability of the partners to raise funds to support their activities. The work of some of the smaller groups has been compromised by inadequate funding, but partnerships with some of the larger and more stable groups appear to be making good progress. Summer schools for educators and outreach professionals would likely accelerate their use of NVO, and the initiative to seek funding for such summer schools is a good one.

Astronomy has already established the values of data sharing and reuse, in large part through the efforts of the NVO, and other fields of research are beginning to explore the applicability of this model to managing their own data sets.

The automated techniques for assessing *compliance* of data sets with metadata standards are an important step forward. Innovative techniques and community involvement may be necessary for assessing the *quality* of the data sets, especially ones that are not part of large-scale, well-documented surveys.

With the advent of electronic detectors, which have enabled the accumulation of massive data sets, the preservation of original data sets for reuse has become an increasingly challenging issue. The Committee was impressed with the initiative proposed by the NVO team to begin to address this issue. Resources beyond the core NVO funding will be required to pursue this effort, but as with the NVO itself, astronomy is in a unique position to provide leadership in exploring the options and feasibility.

The Final Year of the NVO Development Proposal

The meeting of the NVO Advisory Committee occurred at the beginning of the final year of the initial development of this enabling new tool for astronomical research. The Advisory Committee is again extremely impressed by the remarkable progress since our last meeting. The tools that are now available (Registry, DataScope, Open SkyQuery, Spectrum Services, etc.) should enable the community to explore multi-wavelength data sets in ways that were simply not feasible before.

Because this capability is likely to become as essential to the community as, for example, the journals or ADS, it is important to document for the growing NVO user community the accomplishments to date and to enable a smooth transition to the next, more operational, stage of the VO effort. Accordingly, the Advisory Committee recommends that the project, as an important and integral part of its work in the final year, document both the achievements and the vision for the future in a final report for the NSF and the community. Results should include a thorough discussion of lessons learned and documentation of the capabilities developed, plus the basic description of the expected operational model for the NVO, which should include governance and staffing, along with estimates of the computational, storage, and other resources required for the operational era.

As part of its plans for the final year, the NVO team presented a list of potential tasks and asked for advice on priorities. The team itself is better placed than the Advisory Committee to allocate their resources to complete the highest priority development and definition tasks. We do believe, however, that the tasks that the NVO team has identified as remaining to be completed are likely to require both more time and more resources than are available. One of the first actions of the NVO team this year should be to prioritize these tasks, keeping in mind the above requirement for a clear and complete summation of the efforts during the past five years. A reasonable match between tasks and resources should be made, in consultation with the NSF, in order to ensure a smooth transition. The remaining tasks will have to be carried out by the operations team during the next phase of the NVO project. In this context, we also commend the team for its

past efforts to prioritize development tasks to best match the national needs for the National Virtual Observatory.

Recompetition

With the completion of the initial five-year funding for the NVO, the NSF plans to issue an open call for proposals for the next phase, which will include a greater emphasis on operations as well as continued development. The committee recognizes the importance of the competitive peer review process in the awarding of NSF grants, and it is our judgement that, having proved the concept, the current team is well positioned both in terms of expertise and in vision to respond to the call for proposals. However, the committee wishes to emphasize its concern about the impact of significant delays and/or changes in the core NVO team in the upcoming transition to the operational phase. The transition to an operational facility is both challenging and critical to the VO concept, and we hope that the uncertainty associated with the proposal process will not prove too distracting to the team or impact negatively the ongoing development of services for the community. Accordingly, we urge the NSF to conduct its proposal solicitation in a timely manner.

Community Engagement: The Research Community

While the focus of this first NVO phase has rightfully been the development of the VO backbone and a suite of first applications capable of demonstrating the potential for data mining, engagement of the astronomical community in using the NVO and its tools to conduct research and solve real problems is necessarily beginning to gear up. If anything, we have cautioned the NVO team against promising too much to the community, and it appears to us that the team has achieved the proper balance between development and engagement at this point. While the research potential of the VO is still not widely understood, there is a growing community of researchers who are either already engaged in using the VO or are developing plans to do so. Many of these individuals are either active in VO development or have become exposed to it through team members or participants in the NVO School (see below). As the first phase draws to an end, increased effort must go to ensuring community awareness and support of the VO concept.

We encourage continued effort to make available a relatively small suite of well-tested, well-documented, and useful tools that both demonstrate the potential but also provide immediate service to the research community. We suggest it might be useful to provide a short "cookbook," which would contain recipes for some of the likely-most-common tasks. Right now, terminology like SQL/PHP/VOTable/etc. is still mysterious to large segments of the astronomical community, who need some help to get started. A quick set of recipes would be useful. At least one of us has found helpful the SDSS Sky Server document Appendix "A Detailed Narrative of the Twenty Queries" in the technical report by Gray et al. (2002; found at: www.sdss.jhu.edu/ScienceArchive/pubs/msr-tr-2002-01.doc). Giving some examples that demonstrate slightly more than the simplest queries in everyday context (such as how to convert SDSS line widths sigma into km/s, a

more meaningful unit) would quick start the learning curve. Such documents need to be easy to find at the VO page, and also easy to ignore by those who do not need them.

The two (2004 and 2005) NVO schools appear to have been very successful at exposing and training small groups of individuals in VO technologies and potential. The scope of this effort has been appropriate to the development phase, again with the right balance between effort by the NVO team and return on the investment by exporting their expertise to the participants. The distribution of school presentations/notes and demonstration of examples developed by the participants have the potential to expand the knowledgeable community. Kudos to all who have made this program such a success.

A prime target audience for an introduction to VO tools and services might be the summer REU students who participate in programs at the major observatories and departments. Since many of those institutions have resident VO users/developers, we suggest that demonstrations of VO tools might be arranged early in summer programs. Then the undergraduates might even end up teaching their mentors!

We also note that the NVO web site is much improved and makes it much easier for newcomers to explore the capabilities of the NVO. Some sections, like the one on “What is the NVO,” still read rather negatively. The NVO team has accomplished a remarkable amount, and the web site should assert positively what the NVO *can* do.

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. If links to papers that have used the NVO were made available on the NVO web site, potential users might gain more rapid insight into how to make use of the NVO for their own research.

Community Engagement: Education and Public Outreach

Development of a robust education and outreach program requires a well crafted team that includes expertise in website development, programming, content, educational, evaluation, and management. Because the NVO EPO program has less than 1 FTE devoted to this activity, the strategy for NVO EPO has been to partner with education and outreach professionals engaged in existing programs. The risk in this strategy is that implementation depends on the funding available to the partners. Two of the core partnerships (Project Lite at Boston University and Project CLEA at Gettysburg) need additional funding and technical support to complete their NVO projects. Three other partnerships (Virtual Cosmos at UC Berkeley, the Adler Planetarium, and JHU/SDSS) are apparently proceeding well, although we did not hear a detailed update.

The NVO EPO web site was completely redesigned, and both the content and NVO Explorer were refurbished. This is a best effort basis with graphics design contributed

primarily by JHU staff. Redesign included teacher feedback and some other commentary from various individuals. The main result of the evaluation is that the background information is very useful, but the NVO development tools are too difficult to use in a classroom.

In order to address this issue, the NVO EPO program is seeking funding for summer workshops to inform educators and public outreach professionals about the potential of NVO for education and to foster new partnerships. Given the success of the summer workshops for research astronomers, it seems likely that this would be a good strategy for jump-starting the educational use of the NVO tools.

Data Quality and Curation: Requirements

The NVO has made important progress developing standards for data sharing. Applications developed within the project have demonstrated the utility of VOTABLE as a medium for interoperable exchange of table-oriented data among several applications. The advisory committee remains highly impressed with this aspect of the NVO work.

As the project continues, it will become increasingly important to focus on issues relating to the provenance, integrity, and long-term curation of the NVO datasets. It is clear from the presentations that the NVO team has invested considerable effort into establishing quality criteria and "grades" to characterize the integrity of datasets, and we applaud these efforts. However, the emphasis has been on evaluating the lowest levels of data quality—whether a data set exists, is syntactically correct, and operates properly within the NVO environment. These properties can be assessed automatically.

Quality assessment of the measurements or results contained within the data is much more challenging, is more difficult to administer without a formal peer review structure, and almost certainly requires some degree of human intervention. As a guide to defining the goals of quality assessment, it is useful to consider the criteria that a research astronomer applies to determine whether he or she can use someone else's dataset with confidence. Uncertainties in measured quantities must be quantified, with error sources described in the accompanying documentation. The parent data sets, calibrations, and other foundational parameters need to be clearly documented (e.g., for data products from a space mission, the pipeline data version and calibration version used to generate the data should be identified). When data are revised or updated, all changes should be clearly flagged and/or documented, with links back to prior versions of the dataset if possible. Finally, and almost obviously, a user should have confidence that the data will be readily accessible in the future, so those who wish to take advantage of prior work will be able to access the parent datasets and validate the veracity of their results.

This level of curation and documentation has become common practice for large projects and data archives, and this has been critical to the acceptance and usage of archival data by the astronomical community. During the presentations on this topic, however, we became concerned that responsibility for maintaining many NVO datasets would be vested entirely with the individual owners, with no central oversight, monitoring,

validation, or quality assessment after a dataset was initially registered. In such a model, there is nothing to prevent an individual from altering or augmenting the data without notice, or even removing it from the VO. If left unchecked, this will inevitably lead to degradation in the completeness and quality of the aggregate body of data as undocumented changes are made and datasets disappear from the public realm. As the VO grows and evolves such problems will *not* be confined to individual datasets, because the whole premise of the NVO is one in which individual datasets are aggregated and distilled to create new super-sets. Therefore undocumented errors and problems that grow within individual datasets can quickly propagate to others, and only a small subset of "bad apples" could compromise the integrity of a much larger body of resources.

We therefore encourage, during future development of the NVO, that careful consideration be given to two issues: 1) quality assurance and documentation of NVO-registered datasets; and 2) whether archiving in a national- or at least an institutional-level repository with guaranteed long-term access should be made a requirement for high-grade certification of an NVO dataset, if not a requirement for all NVO datasets.

Data Quality

As noted earlier, large projects, including both ground-based surveys and major NASA missions, have undertaken the effort to document, archive, and provide access to their large data sets. The high quality of curation has been critical to the broad acceptance and increasing usage of the archival data by the astronomical community.

The NVO team has consistently argued in favor of open-inclusion of data in the NVO. This means that data sets both large and small will be accessible. We agree that attempts to assess the quality of a large number of data sets prior to inclusion, especially small ones, would be difficult to automate and might well find the NVO staff itself allocating disproportionate fractions of its effort in policing a small subset of the least reliable data.

An innovative approach would be to engage the astronomy community in on-line quality evaluation (and perhaps eventually quality control) of the observational data being provided. A simple mechanism that would solicit and publish comments with respect to specific data sets, and be retained as an essential part of these data sets, is readily feasible. Such a mechanism might prove to be valuable in itself or as a step toward on-line peer review of the data. The example of readers' reviews provided by Amazon.com could provide an initial model. The first implementation might be entirely un-moderated to encourage comments. At a later point it might be necessary to appoint a moderator for some, and perhaps all, of the data sets. The installation of a monitor might eventually generalize to the role of a journal editor and the comments to those of selected peer reviewers.

A considered review of the techniques developed by the commercial on-line data providers, including Google, Yahoo, etc. as well as Amazon, might be instructive. The first step would be to observe the behavior and the features of the on-line user interfaces. A more substantive step would be to reach out to the technical staff at one or more of

these organizations. These organizations are extremely technically competent, and their tasks, goals, and challenges have substantial overlap with the NVO.

There are many other opportunities to use the new technologies in new ways. The NVO has an opportunity, by example, to stimulate the community to develop other innovative astronomical research activities using digital technology.

Data Preservation

The committee was impressed with the efforts being made by members of the VO collaboration to address the problem of archiving and curation of large datasets associated with published journal articles. In years past the published scientific literature contained all of the data used in the generation of the refereed papers. With the advent of modern digital technology that became infeasible, at least in the sense of including large original data sets in the paper version of the journal. Traditional paper-based publication has now essentially reduced data to a second-class citizen, providing only tables or figures in publications from which the original data are rarely recoverable. Unfortunately, this represents a major loss to the continuity of the historical record. This drastic reduction of information bandwidth is unnecessary in the digital environment, and it is now quite feasible to put practical corrections in place.

The AAS journals and *Astronomy and Astrophysics* in particular now have systems in place for publishing and archiving (within the journals themselves or in collaboration with CDS) digital materials including machine-readable tables, extended digital figure sets, animations, source code, and more complex data including FITS image sets. In addition, they have created the mechanisms for direct URL linking to more extensive datasets held at the major astronomical data centers, as well as object tagging via CDS and NED.

The challenge now is to make available the original data from which the summary tables and figures were derived. These data sets may be very large; for example, researchers may wish to have access to the original images. The NVO is designed to enable data sharing, and one important application is the reuse of original data sets. Clearly, however, the reuse opportunities for data are eliminated if attention is not given to the persistence of the data and preservation by trusted agencies. The data need not only to be preserved, but also to be escrowed, preferably in long-lived institutions that may well be separate from the laboratories where the data originated.

Digital preservation is a well-known complex problem in the digital library community. Earlier meetings of the advisory committee suggested that this was an opportunity to work with the research library community, which has increasingly turned its attention to data curation. There is significant interest in the digital library community in new object formats that integrate data and publication text. These new formats effectively allow the reader to use the publication “document” as an entry point to the data set, and as the basis for future research.

At this meeting of the advisory committee, the NVO team presented a plan for moving forward in this area. This plan involves the initiation of a project that investigates the technical, policy, and economic issues of data curation. Technically, the project proposes the creation of a data curation “appliance” that could be easily installed and maintained by libraries. The project also proposes working with two university libraries, Chicago and Cornell, which would install the appliance and experiment with incorporating data curation into their organizational workflows. The NVO is currently seeking funding for this project, and has received seed funding from SPARC (Scholarly Publishing and Academic Resources Coalition).

The advisory committee was quite enthusiastic about this plan. The scaffolding that the NVO has thus far built provides the perfect foundation on which this prototype could be exercised. The results could be of great relevance to the developing e-science, e-research community.

This is a non-trivial effort, and the committee supports the notion of carrying this forward as a separate project with funding from other sources. Because the NVO project is currently in the final year of its initial funding, it would be inadvisable to initiate this program as part of the wrap-up activities of the project. The Board noted also that among the sources for funding this effort, the newly formed cyberinfrastructure office at the NSF should be investigated. The data curation goal falls squarely into their venue. The committee felt that this is an important opportunity for the NSF to leverage already well-spent research money.

The enthusiastic participation of the author community will be critical to the success of this effort, and the experience of the journals suggests that this will only happen if the content expansion is cost effective and imposes minimal time burdens on authors, referees, and the journal/publisher staff. To the extent that the journals are expected to provide support for expanded archiving of original data sets, the NVO project team will need to work closely with the journal editors and their parent scientific societies to make sure that the business model is realistic. Such consultation with the AAS and its journals is built into the current project. However, the full value of this enterprise will only be realized with the participation of all of the major astronomical journals (as has been the case with the ADS), and the proposers should consider the expandability and interoperability of this technology across international lines and to the majority of non-AAS journals in astronomy (with very different business models) as part of their study.

During the presentations it was suggested that once the technology for archiving full sets of supporting data for journal articles is implemented, the inclusion of such data might be imposed as a requirement for publication. Although this is an admirable ideal, it would be a mistake for the team to predicate its system on this comprehensive level of archiving. Imposing such requirements would mark a sea change in our discipline, and would open a Pandora's box of editorial issues with regard to how such requirements would be interpreted and administered. In addition, such requirements could only be implemented after the infrastructure for the archiving itself is in place (e.g., for ground-based observations, little of which is archived today). Any business model for this

project will probably need to incorporate considerable flexibility in the scope and usage level of data archiving/linking, at least for the first years of implementation.

The Future

The challenge for the NVO and for the astronomical community as a whole is not merely to do a better job with zeros and ones than was done previously with paper and photographs. A virtual observatory provides a new mechanism for useful access to vast quantities of astronomical data. It adds no new data; it takes advantage of the fact that the data exist in digital format and that modern digital technology allows manipulation and analysis of data in ways not previously feasible. Thanks in large part to the work already completed by the NVO team, an astronomer can now conveniently access remote data collections, effectively federate multiple data sets, search through enormous volumes of data with an efficiency and effectiveness that could only have been dreamed of until quite recently. Nevertheless, as valuable and productive as these abstractions are, they are obvious extrapolations of traditional activities of astronomers into the digital environment.

Moving astronomy into the digital world opens a large spectrum of potential new modes of interaction between astronomers and their data, between astronomers and the data of others, between different collections of data, between astronomers and their telescopes, etc. The advent of virtual observatories is a key first step in the exploitation of digital technology in furtherance of the science of astronomy.