Towards the National Virtual Observatory

A Report Prepared by the NVO Science Definition Team, April 2002

NVO: The Key Questions and a Synopsis (An Overview Chapter)

The progress in astronomy over the past decade has been breathtaking. Remarkable new discoveries, both from the ground and space, have revolutionized our understanding of the universe and its constituents and have captured the public imagination and advanced the scientific literacy in the United States. Yet this is just a foretaste of the events to come in the new era of information-rich astronomy.

The amount of data in astronomy is growing exponentially, driven mainly by the advances in detector technology across much of the electromagnetic spectrum. The sheer volume of information gathered in astronomy, both from ground and space, doubles every year and a half or so, similar to Moore's law. This increase in data volume is also accompanied by increases in data complexity and data quality. The bulk of the information comes from large, uniform sky surveys over many wavelengths, typically containing many Terabytes of information and which detect literally billions of sources. Synoptic sky surveys that produce many Petabytes of information are imminent.

These remarkable quantitative changes in astronomy will lead to some qualitative changes as well. The very style of observational astronomy is changing, with large-scale, systematic exploration replacing the small sample, piecemeal studies of the past. The inherent limitations in wavelength, area coverage, depth, or resolution of these smaller data sets can thus be overcome. The universe is now being explored in a panchromatic way, over a range of spatial and temporal scales leading towards a more complete and less biased understanding of its constituents, their evolution, and the physical processes governing them. Complex astrophysical phenomena require complex, extensive, and multi-dimensional data for their understanding.

This great richness of information poses substantial technical and methodological challenges ranging from issues of information discovery, data access and manipulation to sophisticated data mining and complex statistical analysis needed for their scientific exploration. Fortunately, the advances in information technology can enable us to fully exploit these massive data sets quickly and efficiently, and they allow us to pose and answer new questions about the Universe. The existing information infrastructure in astronomy is not up to the task, and this implies the need for novel applications of information technology.

The Virtual Observatory concept is the astronomy community's answer to these challenges. It represents an organized, coherent approach to the transition to a new, information-rich astronomy for the 21st century. The National Virtual Observatory (NVO) and its counterparts worldwide will represent a new research environment for astronomy with massive data sets, harnessing the power of information technology and the expertise from applied computer science, statistics, and other fields to advance the progress of astronomy in the era of information abundance. The NVO would federate the currently disjoint set of digital sky surveys, observatory and mission archives, astronomy data and literature services, and it will greatly facilitate the inclusion of the future ones. It would provide powerful tools for the effective and rapid scientific exploration of the resulting massive data sets. It would be also an unprecedented venue for science and technology education and public outreach.

The National Academy of Science (NAS) Astronomy and Astrophysics Survey Committee (AASC) recommended the establishment of the NVO, labeling it the "highest priority small project". This recommendation is in recognition of the scientific opportunities offered by the coming large data sets in astronomy and the benefits from the application of information technologies (McKee, Taylor, et al. 2001). In response to the AASC recommendation, NASA and the NSF established this Science Definition Team (SDT), with the charter to: (1) define and formulate a joint NASA/NSF initiative to pursue the goals of the NVO, and recap the science drivers for such an initiative; (2) describe an overall architecture for the NVO; (3) serve as a liaison to the broader space science and computer science communities for such an initiative; and (4) provide recommendations for proceeding.

This report is based on a series of meetings and teleconferences held between July 2001 and March 2002. Several key inputs were considered: the AASC report; a white paper generated by an *ad hoc* panel and presented to NASA and NSF in May 2000; a proposal to the NSF information technology research (ITR) program; a broader community input; and the proceedings of workshops and meetings held to develop the NVO concept.

As the SDT began its work, a variety of views emerged as to what the NVO is and why it is needed now. At the end of our deliberations, we reached a unanimous consensus endorsing this concept and its timeliness. The SDT used a series of key questions about the NVO to focus the discussion, and these provide a framework for introducing the remaining chapters of this report.

The remainder of this chapter lists these questions and the answers formulated by the SDT. The other chapters in the report provide more detail about the issues expressed here. Chapter 1 summarizes the need for and motivation behind the NVO initiative. Appendix A describes the existing foundations of the NVO. Chapter 2 and Appendix B describe the science drivers and ends with a derived list of science requirements. Chapter 3 and Appendix C describe functional roles and technologies needed for the implementation of the NVO. Chapter 4 addresses the EPO issues. Chapter 5 discusses implementation and management issues. Appendix D summarizes lessons learned from some related programs, and Appendix E provides a draft of a plausible budget for the development of the NVO. Chapter 6 provides our summary and a list of recommendations. Appendices F through H list the team membership, list of references and Web resources, and acronyms used throughout the text.

1. What is the NVO?

As conceived by the National Academy of Science in its Decadal Survey, the NVO will link the archival data sets of space- and ground-based observatories, catalogs of multi-wavelength surveys, and the computational resources necessary to support comparison and cross-correlation among them. It will also provide tools for analysis, visualization and object classification, links to published data, and inclusion of new ones. The NVO will be a complete research environment for astronomy with massive and complex data sets.

2. What will the NVO do?

The National Virtual Observatory will bridge the vast yet separate collections of astronomical data from space and ground-based observatories, providing rapid and seamless access to our knowledge of the Universe. The NVO will embrace selected existing astronomical data from NASA, NSF, international and private observatories, and it will provide a framework for *all* future astronomical data. Even just a few years ago, joining such large "mega-source" databases would have been a daunting task. But today new information technologies enable the public to access dispersed databases as if they were a single entity. These new technologies can allow astronomers to perform a single search across data from a variety of observatories in a variety of wavelength regimes, and to then join these dispersed data together so they are presented seamlessly for further analysis. Information technologies can also unite data from current and planned all sky surveys with existing space- and ground-based archives, minimizing duplication of effort and maximizing discovery potential.

3. Why do we need the NVO now?

A driving force behind the data growth in astronomy is the emergence of sky surveys charting millions to billions of objects in unprecedented depth and measuring tens to hundreds of attributes for each one of them. Effectively combining all these new and existing data sets requires new technologies to provide efficient search, retrieval, and cross-identifications among the archives. Standards must be developed now to avoid wasteful duplication of effort; the longer we wait to retrofit these data, the more expensive and time-consuming the task will become. Some of the existing examples include the SDSS, 2MASS, GSC-1 and -2, DPOSS, NEAT, LONEOS, NVSS, and FIRST. Larger surveys and survey-dedicated telescopes are planned (VST, Vista, CFHT legacy survey, QUEST-2, many asteroid surveys, etc.), culminating in the LSST, a 6.5-meter optical telescope designed to provide deep surveys of the entire sky every few days. In addition, a growing number of ground-based large aperture telescopes (e.g., VLT), now offer Internet-accessible archives, as do most NASA mission and data centers. Future space-based missions such as NGST will deliver orders of magnitude more data. Solar efforts, such as SOHO, TRACE, SDO, GONG+, SOLIS and the ATST are also providing floods of heterogeneous data. Thus it is imperative to act now, otherwise the opportunities provided by combining these large data sets will inevitably be delayed, the costs of combining them will be enormously increased, and the opportunity to apply them to steer the observations of the large observatories will be lost.

4. What new science will come from the NVO?

By providing the tools to assemble and explore massive data sets quickly, the NVO will facilitate and enable a broad range of science. It will make practical studies which otherwise would require so much time and resources that they would be effectively impossible. Federating massive data sets over a broad range of wavelengths, spatial scales, and temporal intervals may be especially fruitful. This will minimize the selection effects that inevitably affect any given observation or survey and will reveal new knowledge that is present in the data but cannot be recognized in any individual data set. NVO-based studies would include systematic explorations of the large-scale structure of the Universe, the structure of our Galaxy, AGN populations in the universe, solar interior structure, variability on a range of time scales, wavelengths, and flux levels, and other, heretofore poorly known portions of the observable parameter space. The NVO will also enable searches for rare, unusual, or even completely new types of astrophysical objects and phenomena. For the first time, we will be able to test the results of massive numerical simulations with equally voluminous and complex data sets. The NVO-enabled studies will span the range from major, key project level efforts to supporting data and sample selection for new, focused studies of interesting types of targets, both for the space-based and major ground-based observatories.

5. How is the NVO different from the archives we have now?

The current data archives are largely disconnected islands. Researchers can search and retrieve data very effectively locally, but there is no coherent service providing cross-archive searches, correlation, or data compatibility. Researchers must search multiple sites (assuming they even know about all the useful sites) and manually join the retrieved data into a single entity. This can often be done for small samples of objects, but it becomes dauntingly inefficient for the millions of objects being generated by the new large sky surveys. Software and network technology has now reached a level of maturity that can support the development of a set of services available through the Internet, and it can now be applied to the field of astronomy. NVO will effect the synthesis of all this.

6. How is the NVO different from past efforts to unite archives?

The NVO is more than just linking archives. The objectives of the NVO are much broader than providing a common user interface to distributed data archives. The NVO will combine data discovery, data retrieval, data comparison, and data correlation tools into an integrated system, providing the necessary computational and data management services to the user automatically. Past efforts to provide some of these capabilities, such as the original Astrophysics Data System, were hindered by having to develop much of the enabling technology from scratch, from the use of proprietary software and tools, and from the imposition of external requirements on internal systems. We now can take advantage of industry-wide IT developments – with increasingly sophisticated facilities for combining and understanding complex, distributed data sets – and apply them to the astronomy domain without having to re-engineer existing systems. We also now have a strong archive infrastructure in place, at least for NASA mission data sets, and more than a decade of additional experience in all aspects of data management and information services.

7. What are the broader scientific benefits of the NVO?

The problems and challenges associated with utilizing large data sets in astronomy will soon emerge in the other areas of science. Scientists in the Statistics and Computer Science disciplines have found astrophysical data to be of particular interest because of its size, complexity and its non-proprietary nature. As a result, the NVO is already establishing substantial partnerships with applied computer science, statistics, and information technology groups, and it will provide a stimulus and a development arena for these fields as well. New technical solutions, algorithms, methodologies, etc., developed in the course of these collaborations will eventually benefit not only other fields of science but also other areas of activity in society and the economy as a whole.

8. What are the societal benefits of the NVO?

The NVO EPO program will bring knowledge of our Universe and the excitement of discovery into the classrooms and homes of America and the world. The NVO EPO effort will be designed to provide the user with an integrated view of the Universe - a system that goes beyond the traditional online portals to the various datasets. The integrated nature of NVO will provide users the opportunity to build knowledge about the Universe by comparing, integrating, and analyzing information from diverse archives, a unique capability provided by NVO that is not offered by individual data archives. In addition to the existing outreach efforts, NVO will serve members of the art, entertainment, and pre-service teacher communities, and it will enhance the role of amateur astronomers as ambassadors of space science and astronomy to the general public. The NVO effort also provides a unique opportunity to enhance technology literacy in a broad sense. The NVO will inform, excite, and educate the public about space science and astronomy, and serve as a catalyst for scientific and technological literacy in the United States.

9. Where will the NVO be located?

The inherent nature of the NVO is that of being geographically distributed. The overall expertise needed for the NVO is broadly distributed across the nation. Moreover, data should be curated by experts and reside where they are located. In view of the very large scale of current datasets, together with their rapid rate of growth and dispersed nature, it is clear that any successful incarnation of the NVO must be distributed in nature. Not only is a distributed structure more efficient, more responsive and more easily implemented, it is also clear that a distributed structure is essential to the success of the NVO. If all the datasets now at hand were to be centrally located, the time required would be such that the current distributed datasets would have doubled in size; thus a central repository would never be complete.

10. How will the NVO be managed?

The NVO presents unique management challenges because of its distributed nature and because it must accommodate funding from multiple sources. This implies that the management structure must be carefully designed and tailored to these needs. Once this structure is in place, it will follow established methods of project management. A work breakdown structure derived from a set of science requirements will be used to drive milestones and deliverables. An NVO project office will direct the implementation and coordinate standards, and an oversight board will monitor the development. The NVO is also a global activity, for it will allow interconnection with parallel efforts now underway internationally (e.g., the European AVO, EGSO, AstroGrid and AstroVirtel projects). Thus, the NVO project will interface with its international counterparts to ensure development of a single set of standards and interfaces.

11. How much funding is needed for NVO?

The NAS Decadal Survey estimated that $70 million is required over a 10-year period to implement and operate the NVO. We suggest a somewhat higher plausible budget of about $90 million (in real year dollars) in new funding over a 10-year period. This total includes funds to join the archives together, to develop tools, and to actually use the NVO for research. A substantial fraction of the funding (approximately 30%) will be provided for research grants and a fellowship program to perform science projects with the NVO. This will be especially important in the early years to demonstrate and drive the NVO development. The level of funding for the grants program should be comparable to that provided to develop the NVO core capabilities. Funding for undergraduate research fellowships should also be provided, as well as support for an overall NVO pre-college Education and Public Outreach program. The NVO will lead to cost savings for future space- and ground-based observatories by providing a set of tools and standards that can be used off the shelf. New archives and observatories will plug into the

NVO, much like new web sites appear on the Internet. A sample budget for the NVO is included in Appendix E.

12. Has any funding already been spent on NVO?

Some funding awarded through existing information technology opportunities is already supporting the development of basic elements of the NVO. The largest is a five year, $10 million effort funded through the NSF Information Technology Research (ITR) program to develop services capable of solving some large scale science problems. Other small and medium NSF ITR programs have been approved for NVO related projects, and several smaller efforts are funded under NASA's AISRP program. In Europe, similar sized efforts are underway, with a total of $15 million committed to four projects: AVO ($3.7 million for FY02-06), AstroGRID ($7.3 million for FY02-06), AstroVirtel ($1 million for FY02-04), and the EGSO ($3 million for FY02-FY04).

13. What are the key steps to implement the NVO?

The NVO must be implemented incrementally and with the widest possible involvement of the astronomical and information technology communities. There are several key steps:

Phase I: Conceptual design, expanded definition of science drivers, implied technical capabilities, general, management, and costing issues; early development work, including further development of prototype NVO services that are funded through the existing grants and programs. CY 2002 - 2003.

Phase II: Definition of the NVO operational/management structure; detailed implementation plan; increased capabilities implemented within the existing data centers, surveys, and observatories; increased community input and involvement; initial development of archives for major ground-based observatories; dedicated NVO science funding. CY 2002 - 2005.

Phase III: Implementation of the full-fledged NVO structure with international connections; commencement of major NVO-based science programs; start of routine operations. From CY 2006 onwards.

14. What are the major challenges in developing the NVO?

There are many technological challenges associated with the NVO in the fields of data storage, data access, data discovery, metadata, standards, interoperability, data-mining, visualization, multivariate statistical analysis, etc. Interdisciplinary partnerships offer paths to solving these challenges.

A major management challenge for the NVO is to coordinate a highly distributed effort, embracing a single set of goals to integrate national and international ground- and space-based archives. Within the United States, the major funding will come from the NSF and NASA, which will require the two agencies to jointly manage the activity, decide their relative roles, and provide the appropriate level of funding. We expect the new joint NASA-NSF National Astronomy Committee will help address this issue. A challenge here is to ensure that the NVO produces a value-added product that clearly delivers major benefits to the science community. This will require both a further development of science requirements and a clearly focused approach in implementing the NVO.

15. How will we know that the NVO is a success?

The NVO metric of success is how much it will increase the science productivity of all astronomers. With the NVO, projects that today might take months or years will be achieved in minutes or hours. The successful NVO will, in its mature form, be seen as an essential element in both astronomical research and in education and public outreach. NVO activity will be at the cutting edge in advancing research capability and will be at the focus of prominent research ventures. 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. Periodically, NASA and NSF advisory groups should review the NVO activity to ensure that it is fulfilling its mandate and that it is responding to the changing needs of its user communities.

Selected Bibliography & Web Resources

Banday, A., *et al.* (editors) 2001, *Mining the Sky*, A. ESO Astrophysics Symposia, Berlin: Springer Verlag.

Brunner, R.J., Djorgovski, S.G., & Szalay, A.S. (editors) 2001, *Virtual Observatories of the Future,* Astronomical Society of the Pacific, Volume 225.

McKee, C., Taylor, J., *et al.* 2000, *Astronomy and Astrophysics in the New Millennium (Decadal Survey)*, National Academy of Science, Astronomy and Astrophysics Survey Committee, Washington D.C., National Academy Press. Also available online at http://www.nap.edu/books/0309070317/html/.

NVO White Paper, in Brunner, R.J., Djorgovski, S.G., & Szalay, A.S. (editors) 2001, *Virtual Observatories of the Future,* Astronomical Society of the Pacific, 225, 353. Also available online at http://www.arxiv.org/abs/astro-ph/0108115.

Szalay, A.S., and Gray, J. 2001, *Science*, 293, 2037.

The NVO SDT Web Page: http://www.nvosdt.org

The US NVO ITR Project Web Page: http://us-vo.org

The VO Forum: http://voforum.org/

The European AVO Project Web Page: http://www.eso.org/projects/avo/

The UK Astrogrid Web Page: http://www.astrogrid.ac.uk/