White Paper:
A National Virtual Observatory
for
Data Exploration and Discovery

October 21, 1999

1 Towards a National Virtual Observatory

A major paradigm shift is taking place in astronomy. In the past, the majority of astronomical observing consisted of individual observations of a small sample of objects, usually over a narrow wavelength range. The Palomar Sky Survey plates served as a reference catalog of the sky, with visual inspection as the access method.

The situation is changing dramatically. New dedicated survey instruments, such as Sloan and 2MASS, are producing uniform, high-quality photometric and spectroscopic observations of millions of objects and these observations will be available in digital form on the desktop of every astronomer. The science returns from these surveys will be enormous: from the discovery of large numbers of high-Z quasars, to the characterization of the moderate-redshift universe, to the structure of our own Galaxy, to definitive studies of low mass stars and brown dwarfs.

The new surveys are providing tens of Terabytes of catalog data for the astronomical community and will be providing Petabytes of data by the end of the next decade, covering many different wavelengths and many different epochs. The volume and quality of data from these surveys will be increasing every year, and there will be an enormous pressure from the astronomical community to integrate the separate archives into a seamlessly

inter-operating entity which will allow true multi-wavelength astronomy to be performed on entire classes of objects.

In addition, astronomers will need integrated access to the archives of individual observations from ground and space-based observatories, seamless connection to information archives such as ADS, NED, and Simbad, as well as legacy data archives. We recommend the creation of a *National Virtual Observatory (NVO)*, to accomplish the integration of the Nation's valuable astronomical data. Investing resources on the integration of the existing and forthcoming high value astronomical data should be seen as a priority for the coming decade.

A few years ago such an endeavor would not have been technically possible. Recent developments in database technology, high-speed networking, and storage systems make a National Virtual Observatory not only feasible, but critical for exploiting the next generation of astronomical data. An NVO would be a true national observatory facility: it would be general purpose and capable of supporting a large number and range of astronomical investigations initiated by individual researchers. It would give astronomers multi-wavelength access to large areas of the sky. However, it would be unlike any current observatory. The NVO would be a virtual observatory, supporting astronomical observations and investigations via digital representations of the sky and associated electronic information archives. It would be inherently distributed in nature, using the next generation of high-speed networks as its infrastructure.

An NVO, like other observatories, would enable a wide range of unique and important astronomical research: multi-wavelength identification of large candidate samples of objects such as brown dwarfs, high-z quasars, gravitational lenses, and ultra-luminous IR galaxies; multi-wavelength cross-identification of sources discovered in new surveys and observations; and searches for rare and exotic new objects among the billion or so catalogued sources. An NVO would be an important new tool for several main-line fields of astronomical research: large-scale structure, galaxy evolution, active galaxies, galaxy clusters, galactic structure, and stellar populations. We anticipate that the NVO would become an indispensable research tool for the entire astronomical community.

Most importantly, an NVO would provide the basis for enabling science of a *qualitatively different nature*. Whereas in the past, exploratory efforts were constrained by the need to carefully select small samples of objects, we

can now imagine exploiting the revolution in computing and networking to carry out new and different type of astronomical research: multi-wavelength exploration and discovery over the entire sky using all known catalogued astronomical objects. The exploration process will include: discovery and identification of unique astronomical objects through unusual colors; discovery of patterns revealed from the analysis of statistically rich and unbiased image and catalogue data; and, gaining new insights into complex astrophysical systems through confrontation between sophisticated numerical simulations and the data. The discovery process will be accelerated through the application of advanced visualization, data mining, and statistical tools.

Why do we propose this now, and why do we think that it would succeed? Similar integration efforts in other areas of science have not always been successful. But, it is important to note, that astronomers (and space scientists) form a relatively small community, compared to the medical or earth science fields, for example. Astronomers have been able to create and adhere to standards in the past, like the well established FITS format, that cuts across many sub-disciplines of astronomy. Based upon standards, we have developed and used portable tool that make collaborations much easier. Perhaps the most important reason why such a unification or merging of data sets should be successful is that we all look at the same objects on the sky, and the timescale of changes is relatively slow for most objects. The number of pixels on the sky is mostly determined by the atmosphere (approximately 1 Terapixels over the whole sky), thus the different data sets in the different bands are about the same size and complexity. The time is right, and the data are just becoming available.

1.1 The New Generation of Survey Science

There are numerous large digital survey projects under way, which together will create a unified multi-wavelength view of the Universe. Efforts include 2MASS, SDSS, DPOSS, FIRST, NVSS, COBE, MAP, GSC-II, ROSAT, GALEX, and GLAST, with wavelengths from γ-rays and X-rays through UV, optical, near-IR to microwave and radio. In addition, the use of major ground-based and space-based observatories for survey work is increasing, with NOAO now allocating 20% of its observing time to surveys, and the NASA Great Observatories providing opportunities for large survey projects. The catalogs of these projects will provide all-sky (or almost all-sky) informa-

tion in more than 15 different bands. Their main characteristics are that they have been designed to be homogeneous, are well calibrated, and have good control over systematics. The catalogs will typically contain over 100 million objects each, and will be multiple Terabytes in size. Once these catalogs are completed, their combination will allow unique science. The overlap between them will be substantial, allowing broad-band spectra to be constructed for individual objects, but at the same time there will be a lot of non-detections between different surveys indicating unusual colors or transient objects.

Scientific analysis of digital astronomical databases has already yielded exciting results. We give a few examples:

- Study of the lower main sequence using 2MASS leading to a new spectral class of stars, L Dwarfs
- Detection of numerous Brown Dwarfs using 2MASS and determination of the Brown Dwarf mass function
- Detection of methane dwarfs using SDSS
- Identification of dozens of high-z quasars using DPOSS, FIRST, and SDSS
- Analysis of variable and binary stars using MACHO and OGLE

Some of this work is primary, in the sense that the database was created with that work in mind. Examples of this mentioned above include the detection of brown dwarfs and study of the lower main sequence. Some are secondary, and involve analysis of databases for purposes not envisioned by the database creators. Examples include variable and binary star analysis using the MACHO and OGLE data which were designed for the detection of gravitational micro-lensing. Some result from intercomparison between different databases. Examples include intercomparison of radio and optical data to detect high-Z quasars and the use of ROSAT and MACHO databases to study supersoft X-ray binaries.

Exploiting the complementarity of surveys will become increasingly important. Surveys at different wavelengths provide views of rest frame galaxies at different redshift epochs allowing the study of their evolution. Some surveys concentrate on angular resolution and astrometry while other concentrate on wavelength coverage and spectroscopy. Combination of the surveys will significantly increase the value of each individual survey.

We expect several flavors of scientific results to come from the all-sky multi-wavelength studies enabled by an NVO:

- **Enabling Science.** Somewhat mundane, but perhaps most important, is the enabling science that NVO will provide. Cross-comparison of objects at wavelengths from radio to gamma-ray will allow observers to efficiently construct lists of target objects which can then be followed up by detailed spectroscopic and imaging observations. A good example is the combined use of radio and optical data to select candidate high-z quasars. We anticipate that an NVO will become the premier tool of observers in efficiently designing observing programs for large ground-based and space-based telescopes.

- **Population Studies of Galactic and Extragalactic Objects.** Unbiased surveys of astronomical objects are essential for determining luminosity functions, mass functions, and evolutionary characteristics. The NVO will allow access to catalogs of billions of well-calibrated objects with high-quality photometry and spectroscopy. Research areas such as large-scale structure, Galactic structure, and galaxy evolution will benefit significantly from an NVO.

- **Detection of Unique Astronomical Objects.** Identification of rare and unusual astronomical objects will be a prime objective of NVO. Detection of such sources is accomplished by looking for objects with unusual "colors". Good examples are searches for Brown Dwarfs and very-red AGN. With colors spanning the range from radio to gamma-ray, an NVO will be able to search for a wide-range of exotic objects. With more than a billion cataloged sources, even objects that are rare at the level of one in a hundred million will occur several times in the data set. Data-mining will reveal these in the data.

1.2 The Next Generation of Surveys

The current generation of planned surveys are generating tens of Terabytes of image and catalog data, at a rate of a few tens of Gigabytes per night. However, this only represents the initial rumblings of the data avalanche to come. Since 1970, CCD capabilities have been on a Moore's law growth curve, while the growth of glass area at major observatories is on a much slower

growth curve. The exponential growth in the pixel count for large format CCDs is driving the revolutionary growth in survey data. The increase in CCD capability has motivated the push to larger fields of view and vastly enhanced multi-object spectroscopy. The growth will continue with Gigapixel arrays on the horizon.

Several new survey projects have been proposed or are being studied. Representative of the type of surveys that are likely to occur by the end of the next decade are the:

Dark matter telescope: map the distribution of matter in the universe for z $<$ 1.5 from analysis of weak lensing revealed through deep, high quality images of galaxies

find moving (asteroids; KBOs; high proper motion stars) and variable (SN; AGNs; variable stars) through repetitive surveys

1 TB per night of compressed data

Wide-Field Spectroscopic Telescope: trace evolution of galaxies from z $\sim$ 4 to the present from analysis of star formation rates; chemical abundances and kinematics

4x16Kx16K CCD Mosaics; 4x8Kx8K InSb Mosaics
=>1.3 Gpixels

These next generation surveys will significantly increase the amount of astronomical data available to the researcher. The number of catalogued galaxies will rise from a few hundred million to 10 billion or larger. In addition to the drive to fainter magnitudes, the new surveys will explore new regimes of time variability, significantly extending the information gathered by surveys such as MACHO and OGLE. The exploration of new parameter space in sensitivity and time variability will generate Terabytes of data per day by the end of the next decade.

1.3 Technical Challenges

The current generation of surveys will revolutionize the field of astronomy, increasing the amount of information available to researchers by several orders of magnitude. Essentially these projects are the astronomical equivalent

of the Human Genome Project or the Earth Science database. The resultant astronomical archives will be large (each exceeding several Terabytes) and complex: textual information, derived parameters, multi-band images, and spectral information. Consequently, there are rather complex technical issues involved in how such amounts of data can be *efficiently served* to the public. Part of the difficulty is that very little prior experience is available on how to manage scientific data on this scale for a whole community. The next generation of surveys will likely increase the sum total of astronomical data by an additional one or two orders of magnitude.

It is important to understand, that today's approaches of accessing astronomical data do not scale well into the Terabyte regime - brute force does not work! Current state-of-the-art hardware cannot support hundreds of 'brute force' queries per day each involving Terabytes of data. Thus intelligent software solutions are absolutely required. With Terabyte catalogs, even customized data sets of a few percent are still in the 10 GB range, thus an intelligent, high level data management is still needed. The next generation of internet technology is aiming at sustained transfer speeds of 100 Mbytes/sec. The astronomical community must prepare itself to take advantage of these future capabilities to insure that its future data analysis needs are met.

The inter-operability of the large data archives poses non-trivial problems. Since most of these archiving projects are just about to begin, it is not too late to coordinate their architectures, but action must be taken very shortly. The individual catalogs are likely to remain at their respective data centers. Their seamless inter-operability will involve common high performance protocols and data interchange formats. Their efficient integration will require exchange servers which will coordinate searches involving multiple archives.

The time to implement such an integration is now, or in the very near future, since in five years many of the surveys mentioned above will be close to completion. After each of the surveys complete their own archives, it will be much more difficult. The archives will have to be rebuilt almost from scratch, or integration will not happen.

2 The NVO Project

The National Virtual Observatory would consist of the main repositories of astronomical data in the US, together with significant computational resources, all interconnected by high-speed networks. It would include information services such as ADS and NED, and compute servers such as those at the NSF and NASA supercomputing sites. It would be responsible for serving data sets to the whole community in a transparent and efficient manner. NVO participants would collaboratively manage the available resources and make sure that data sets are up-to-date and on-line. NVO would also provide links to astronomical name servers and legacy data. Searches for astronomical sources would be executed across the data sets in a fashion transparent to the users.

2.1 Major Elements of the Virtual Observatory

Existing data archives. We envisage that data would optimally reside with the respective groups, who know their own data best. This of course means that the data is scattered across the continent. These groups would receive reasonable funds to maintain their own data. NSF supported data centers (e.g. NOAO) would obtain their funding from the NSF, while NASA funded centers (e.g. STScI, IPAC, HEASARC) would get their funding through NASA. There would also need to be support for continued access to important archival data sets after the active mission phase.

Support for users: query and computation support. There are different types of users, whose support requires a widely diverse array of resources. Much of the general astronomy public will only use these archives on a causal lookup basis, and will utilize a www interface, generating a large number of rather simple queries. These can be relatively easily supported via a central web-site with a state of the art, but not too complex query engine. Intermediate users will want to use the archive in a more elaborate fashion. Their usage should be free, but limited in scope, in order to manage the available resources.

The greatest difficulty arises from the requirement to serve "power users", who will undertake multiple searches through Terabytes of data, extracting

hundreds of Gigabytes for further processing. This task resembles accessing supercomputer resources, and our straw-man idea is to handle it in a similar manner - these searches could be done at one of the NSF- or NASA-supported supercomputer sites. Database time would be allocated in a fashion analagous to supercomputer CPU time. Centers would get funding to support these database activities. As computing technology and networks evolve, more of the computing activity could be accomplished local to the user, perhaps on so-called "commodity supercomputers" such as Linux clusters.

Standards development. A critical and immediate activity of the NVO will be the development of standards for information interchange. Specific areas for standardization include those for: meta-data, meta-services, streaming formats, object relationships, and objects attributes (e.g. position, flux, band).

Education. The NVO will be a wonderful resource for education at all levels, including K-12. Planetariums and public science museums could utilize the resources of the NVO in exhibits and presentations.

We also anticipate benefits for graduate astronomy education. The intellectual work of the National Virtual Observatory will involve astronomers, computer scientists, statisticians, and even mathematicians, providing beneficial consequences for graduate education. For instance, an astronomy graduate student would have some professional training in, say, statistics, and thus have access to a broader range of career options upon graduation. Similarly, a statistics graduate student might perform thesis work mining astronomical data, and upon graduation take a job in industry that exploits the statistics expertise (this has happened already in the SDSS, MACHO and TAOS projects).

Interconnection of sites. The interconnection of the different archive/service sites could be accomplished via one of the testbed programs connected to the next generation of high-speed networks. Extensive use of data mirroring and data caching would also be employed to allow efficient high-speed access to the data sets.

The technology for the high-speed connectivity is here, but there is a lack of credible academic applications, which could generate a realistic network load. The National Virtual Observatory could thus become a prime example of creative network usage.

Information Technology and Computer Science. The NVO requires original development of new tools, which in turn will require the use of state-of-the-art techniques from Computer Science. and will require solution of challenging problems that have broad applications both to other disciplines and commercially. Among the most critical are:

- creating appropriate data structures and storing the data on physical media in a fashion that anticipates and adjusts dynamically to subsequent queries
- carrying out computations on complete multi-terabyte database. Implicitly, these require huge I/O bandwidth and massive computational power.
- developing querying engines that enable users to formulate search and recognition sophisticated queries efficiently
- exploring and possibly developing new languages that integrate query capabilities with data servers
- developing a statistical "toolbox" designed to enable searches for multi-dimensional correlations and confrontation between databases and simulations
- developing a visualization "toolbox" designed to display complex, multi-dimensional data in ways that facilitate discovery of unexpected patterns

The results from NVO will be very relevant beyond astronomy; modern society in general is struggling with the information avalanche. The technology to access and mine the data can best be developed by a wide collaboration that involves not only astronomers, but computer scientists,statisticians, and participants from the IT industry. The NVO would be a very credible interdisciplinary project, which could be funded out of funds allocated

for long-range Information Technology research. Some aspects of advanced data-mining activity could also draw industry support. However, to meet the needs of astronomical science, NSF and NASA must insure that the core support for an NVO be a priority.

2.2 Project Size, Scope, and Timeline

In our view, the variety of skills, their location in multiple institutions with different cultures, and the need for agile deployment of resources to match skills to new opportunities requires a *distributed structure* for NVO comprising:

- a small core group having the skill mix to work with the community to plan and manage NVO activities
- project teams located at a number of institutions throughout the world, each charged with working on one or more of the key problems identified above
- project teams at universities, the national data centers, and the national laboratories each charged with working with NVO to develop protocols, standards and quality assurance for carrying out surveys and populating archives

The NVO will require resources comparable to those of a satellite mission to be successful. In this section, we make a first attempt to scope the size of an NVO effort. We limit ourselves to the scope and cost of the new elements of the NVO, and will not attempt to summarize the total cost of maintaining the National astronomy data resources, which includes the individual surveys and data archives.

The following components of NVO will need support:

- System development: database technologies, query estimation and optimization, standards development, network technology.
- Maintenance of NVO-specific databases (e.g. cross-identification information and custom data subsets).

- User tools and interfaces: applications environment and tool kit, query languages, statistics tool kit, visualization tool kit.
- User services: documentation, user feedback, bug-fixing.
- Advanced research: automatic search algorithms, data-mining, statistics, visualization.

All of the products of the National Virtual Observatory will be kept in the public domain, and will be made accessible to the scientific community as they are created. Data which are incorporated into the NVO will be made accessible as soon as possible after receipt. Analysis and data management tools will also be placed in the public domain as soon as they are created. This latter goal places significant demands on the code development efforts, which will be performed according to professional standards regarding version control, transparency, and documentation.

Several small efforts are already underway in the above areas that could feed directly into the NVO. Seed funding is needed immediately (year 0) to allow planning and initial design of the system. Proposals for participation in the NVO would be solicited during this preliminary year. The support should ramp up quickly during the first year of the actual program to a standing level of at least \$5M/yr and be sustained for 3 years. The 5th year would be a transition year during which the program would be critiqued for scientific productivity and potential continuation. We estimate the total cost of the program, including seed funding, to be a minimum of about \$25M. We note that this cost estimate includes only the funding of that part of the NVO which involves software development for the integration of existing data and information resources. As previously mentioned, significant funding is also needed for the individual data archives, for astronomical information services, and for the networking and computing infrastructure.

We suggest that the NVO fund a range of R&D activities including smaller research grants to individual groups as well as larger efforts at universities and national centers. Mathematics and computer science participation should be encouraged in these efforts.

3 Conclusions

- A National Virtual Observatory would be a cornerstone of astronomical research during the next decade, enabling a wide range of important science.
- The creation of the National Virtual Observatory is critical, given the ongoing paradigm shift in the way astronomical research is performed. Creation of a virtual observatory is unavoidable, the only question is whether it will take place in a planned or chaotic fashion, and whether sufficient support will be available for its proper design, operation, and evolution.
- The NVO can only materialize from a community-wide effort, and only if it is genuinely aimed at serving the whole community. It must have a broad base of participation.
- The necessary technology for the NVO is available or within reach.
- The organizational framework of the National Virtual Observatory could foster multi-agency support and encourage interdisciplinary funding.