Towards the National Virtual Observatory

**A Report Prepared by the
National Virtual Observatory
Science Definition Team**

April 2002

Preface

This Report represents a result of deliberations of the National Virtual Observatory Science Definition Team (NVO SDT), from July 2001 to March 2002. The creation of the NVO was recommended by the NAS Decadal Survey, "Astronomy and Astrophysics in the New Millennium". The SDT was chartered jointly by the NSF and NASA in response to this recommendation, 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.

Following the Executive Summary, the overview chapter (NVO: The Key Questions and a Synopsis) provides a stand-alone overview of the key issues in Question/Answer format. It is followed by the main body of the Report, including several Appendices, which offer more details and supporting materials.

This Report and a number of relevant links are also provided on the SDT website, http://nvosdt.org

Table Of Contents

Executive Summary

Astronomy has become an enormously data-rich science. The cumulative data volume, now measured in hundreds of Terabytes, is growing exponentially, with increases in data complexity and quality as well. The universe is now being explored systematically, in a panchromatic way, over a range of spatial and temporal scales that lead to a more complete and less biased understanding of its constituents, their evolution, their origins, and the physical processes governing them. Complex multi-dimensional astrophysical phenomena require complex multi-dimensional data for their understanding. The inherent limitations of individual data sets can be overcome by combining them and thus uncovering a new knowledge that cannot be gained from any one of them individually. This great richness of information poses substantial technical challenges, ranging from data access and manipulation to sophisticated data mining and statistical analysis needed for their scientific exploration. Our current ability to fully exploit scientifically this data avalanche is limited by the existing tools and resources, and the problem is growing rapidly.

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 would represent *a new research environment for astronomy with massive data sets,* harnessing the power of the information technology and the expertise from applied computer science, statistics, and other fields to advance the progress of astronomy. This novel scientific organization will be a geographically distributed, web-based, open and inclusive environment, with a broad constituency of users and contributors that transcends the traditional divisions of wavelengths, ground vs. space based, etc.

The NVO will 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 future ones. It will provide powerful tools for the effective and rapid scientific exploration of massive data sets. It will be *technology enabled, but science driven*. It will generate more efficient and cost-effective science, and even completely new science, by making practical those studies that today are too costly in terms of the efforts and resources required. It will empower scientists and students everywhere to do first-rate astronomy, and it will become an unprecedented venue for science and technology education and for public outreach.

The scientific and operational requirements of the NVO map into a set of technical capabilities needed in order to fulfill this vision. These include development of the following:

- Standards for accessing large astronomical data sets that can accommodate the full range of wavelengths, observational techniques, and application-specific needs for all types of astronomical data: catalogs, images, spectra, visibilities, and time series; including also standards for metadata, data formats, query language, data and service discovery, etc.

- Mechanisms for the federation of massive, distributed data sets, regardless of the wavelengths, resolution, and types of data (e.g., images, spectra, catalogs, annotations, numerical simulations, etc.).

- Provide new, effective mechanisms for publishing massive data sets and data products, including quality control and links to other published data and to the literature.

- Data analysis toolkits and services, including source extraction, parameter measurements, and classification from image data bases; data mining for image, spectra, and catalog domains, together with federations thereof; multivariate statistical tools; and multi-dimensional visualization techniques for novel and complex data sets.

- System wide gateways to provide access to these capabilities.

- EPO-oriented data access and capabilities.

These capabilities would build on existing foundations and would greatly expand their functionality. Their implementation would lead to a fully operating NVO. We envision and recommend three phases of development:

1. 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.

2. 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.

3. 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.

A preliminary analysis suggests that the total anticipated budget for the period of 10 years will be $ 90 M (in real year dollars) of new funds, about 30% of which will be in grants and fellowships to the community. We believe that an investment in the NVO will result in significant cost savings for future surveys, missions, and other data intensive projects. We recommend the following immediate action items:

A. Definition of a suitable organizational/management structure for the NVO that can accommodate the needs and constraints of both the NSF and NASA and be consistent with the vision, scope, and capabilities needed.

B. Pending the existence of the NVO as an organizational entity, creation of an advisory, planning, and coordinating body which would continue and expand the functions started by this Team.

C. Identification of the funding venues for the development and prototyping of NVO technical, scientific, and EPO demonstration projects, together with steps towards a more dedicated and extensive future funding.

In conclusion, it is the unanimous view of this Science Definition Team that the National Virtual Observatory is essential to the future well-being and competitiveness of U.S. Astronomy, and that this initiative should be fully funded as expeditiously as possible.

NVO: The Key Questions and a Synopsis

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.

1 Introduction: The Birth of a Virtual Observatory

Astronomy, like most sciences, has become immensely data-rich, with data sets measured in many Terabytes, and soon Petabytes. The sky is being surveyed systematically at many wavelengths, with billions of stars, galaxies, quasars, and other objects detected and measured with an unprecedented level of detail. These massive data sets are a new empirical foundation for the astronomy of the 21st century, hopefully leading to a new golden era of discovery.

Yet, the progress in the scientific exploitation of these vast new data sets has not kept pace with the exponential growth in the amount and quality of the information available. We need a new approach, methodology, and infrastructure to fully exploit scientifically these great quantities of information. These data sets and the technology to explore them will produce a qualitative change in the way astronomy is done, and they will open completely new fields of astronomy. What is at stake is nothing less than the ways in which astronomy will be done in the era of information abundance.

The response of the astronomical community to the challenges and opportunities given by the exploitation of massive data sets is the concept of a Virtual Observatory (VO). Within the U.S., this was articulated by the NAS Astronomy and Astrophysics Decadal Survey Report (McKee, Taylor, et al. 2001), which recommended as the highest priority in the "small" (cost < $ 100M) category the creation of the **National Virtual Observatory** (NVO). Parallel VO definition and establishment efforts now exist worldwide.

The NVO will challenge the astronomical community, and yet it will provide new opportunities for scientific discovery that were unimaginable just a few years ago. It is a development that is fundamentally different from traditional advances, such as the building of large new telescopes or space missions, but with an overall long-term impact larger than that of any given observatory. This document describes the scientific opportunities and technical challenges of an NVO, and suggests an implementation strategy aimed at realizing the goals of the NVO in cost-effective manner. It should be viewed as a start of a building process.

1.1 The Need and the Opportunity: Astronomy in the Era of Information Abundance

For hundreds of years, the usual mode of carrying out astronomical research has been that of a single astronomer or small group of astronomers performing observations of a small number of objects. In the past, entire careers have been spent in the acquisition of enough data to barely enable statistically significant conclusions to be drawn. Moreover, because observing time with the most powerful facilities is very limited, many astrophysical questions that require a large amount of data for their resolution simply could not be addressed.

This approach is now undergoing a dramatic and very rapid change. The transformation is being driven by the unprecedented technological developments over the last decade. The major areas of change upon which this revolution in astronomy rests are advances in telescope design and fabrication, the development of large-scale detector arrays, the exponential growth of computing capability, an increasing capability of space-based observatories and missions, and the ever-expanding coverage and capacity of communications networks.

The steep increase in the volume and complexity of available information is based on the great progress in technology, including digital imaging (the chief data source in astronomy), and, of course, the ways of processing, storing, and accessing information. Most of the scientific measurements and data obtained today are either generated in a digital form or promptly converted to one. Essentially all astronomical measurements today are digital in nature, and most instruments contain some form of a digital imaging arrays. Such devices, in turn, are based on the same technology (integrated circuits and microelectronics), governed by Moore's law and are thus growing exponentially in their information-generating ability (see Figure 1).
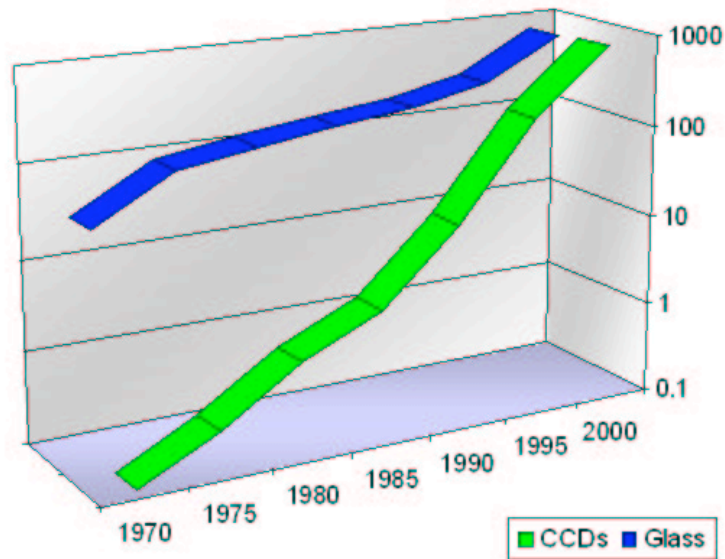
Figure 1. The total area of astronomical telescopes in square meters (upper curve), and CCDs measured in Gigapixels (lower curve), over the last 25 years. The number of pixels and the quantity of data collected doubles every 18 months, following Moore's law.

In addition to this increased data rate, the manner in which observations are being made is also changing. Although the new observatories in space and on the ground still devote a significant fraction of their time to research in the single observer/single program mode where small blocks of time are allocated to many specifically targeted research programs, more time is now being devoted to large scale surveys of the sky, often at multiple wavelengths, that involve large numbers of collaborators. These large survey programs will produce coherent blocks of data obtained with uniform standards and with the amount of data often measured in terabytes. These trends will continue. For example, a facility recently recommended for construction by the AASC decadal survey, the Large-Aperture Synoptic Survey Telescope (LSST) could produce up to 10 terabytes per day!

A major technological development that will change the character of astronomical research is the advent of high-speed information transfer networks with broad coverage. Although the rapid transfer of large amounts of data over common networks is currently unacceptably slow (over 20 days to transfer a 1 Terabyte data set), future networks will be much faster. The availability of

these data rates, together with the high efficiency of data acquisition at both ground and space based facilities, will make possible the efficient transmission of large amounts of data to many different sites. This technology will also enable access to specific subsets of data by an extensive user community that prior to this had no readily available access to these data; the potential scientific yield resulting from this accessibility will be enormous.

These technological developments have converged in the last few years, and they will completely alter the manner in which most observational astronomy is carried out. These changes are inevitable and irreversible, and they will have dramatic effects on the sociology of astronomy itself. Moreover, there is a growing awareness, both in this country and abroad, that the acquisition, organization, analysis and dissemination of scientific data are essential elements to a continuing robust growth of science and technology. These factors make it imperative to provide a structure that will enable the most efficient and effective synthesis of these technological capabilities. Hence there is a need now for an entity such as a National Virtual Observatory to oversee the disposition of the growing body of astronomical data.

As we look ahead, the astronomical community stands poised to take advantage of the breathtaking advances in computational speed, storage media and detector technology in two ways: (1) by carrying out new generation surveys spanning a wide range of wavelengths and optimized to exploit these advances fully; and (2) by developing the software tools to enable discovery of new patterns in the multi-Terabyte (and later Petabyte) databases that represent the legacies of these surveys. In combination, new generation surveys and software tools can provide the basis for enabling science of a qualitatively different nature.

1.2 The Changing Nature of Observational Astronomy

While the quantity of data continues to grow exponentially, our knowledge and understanding of the world has not kept pace. Transforming vast masses of bits into a refined knowledge and understanding of the universe is a highly complex task. The great quantitative change in the amount and complexity of available scientific information should lead to a qualitative change in the way we do science.

Large digital sky surveys and archives now becoming the principal sources of data for astronomy. Increasingly, the field is being dominated by the analysis of large, uniform sky surveys, sampling millions or billions of sources, and providing tens or hundreds of measured attributes for each of them. There is a paradigm shift in observational astronomy, with survey-based science becoming an ever more important way of exploring the universe in a systematic way. The sky is now being surveyed over a full range of wavelengths, giving us, at least in principle, a panchromatic and less biased view of the universe.

We now have the tools to carry out surveys over nearly the entire electromagnetic spectrum on a variety of spatial scales and over multiple epochs, all with well-defined selection criteria and well-understood limits. The ability to create panchromatic images, and in some cases digital movies of the universe, provide unprecedented opportunities for discovering new phenomena and patterns that can fundamentally alter our understanding. In the past, a panchromatic view of the same region of sky at optical and radio wavelengths led to the discovery of quasars. The availability of infrared data led to the discovery of obscured active galactic nuclei and star-

forming regions unsuspected from visible images. Repeated images of the sky have led to the discovery of transient phenomena - supernovae, and more recently, micro-lensing events - as well as a deeper understanding of synoptic phenomena. The joining together of various large-scale digital surveys will make possible new explorations of parameter space, such as the low surface brightness universe at all wavelengths.

Many astronomical surveys, large telescopes, and space missions are already producing large quantities of high quality legacy data, and much of this is currently being archived (a glaring exception are many ground-based observatories, but plans exist to bring their archives on line as well). Most of these data, obtained through use of costly and highly oversubscribed, state of the art facilities, have an unprecedented richness and depth, and they offer unique opportunities for application to a variety of scientific programs by a wide range of users. The existence of such information-rich archives, containing multi-wavelength data on hundreds of millions of objects, will clearly create a demand within the astronomical community for access to the archives and for the tools necessary to analyze the data they contain. Opportunities for data mining, for sophisticated pattern recognition, for large scale statistical cross correlations, and for the discovery of rare objects and temporal variations all become apparent. For solar astronomy, where there is a rich heritage of research using disparate data sets, enormous new data sets are about to become available which will require the NVO to fully realize their science productivity.

In addition, for the first time in the history of astronomy, such data sets will allow meaningful comparisons to be made between sophisticated numerical simulations and statistically complete multivariate bodies of data. The rapid growth of high speed and widely distributed networks means that all of these scientific endeavors will be made available to the community of astronomers throughout the U.S. and in other countries.

The potential for scientific discovery afforded by these new surveys is enormous. Entirely new and unexpected scientific results of major significance will emerge from the combined use of the resulting datasets, science that would not be possible from such sets used singly. However, their large size and complexity require tools and structures to discover the complex phenomena encoded within them. The NVO will meet these needs through the coordination of diverse efforts already in existence as well as providing focus for the development of capabilities that do not yet exist. The NVO will act as an enabling and coordinating entity to foster the development of tools, protocols, and collaborations necessary to realize the full scientific potential of astronomical databases in the coming decade.

1.3 Summary of the Virtual Observatory Concept

The NVO thus represents a response to the scientific and technological challenges and opportunities posed by the massive data sets in astronomy, and it enables their full scientific utilization. It is a mechanism for an orderly and coherent transition to the new, information-rich astronomy.

It represents an organizing framework for the broad spectrum of distributed efforts in order to minimize a wasteful duplication of work and resources and to create useful scientific functionalities quickly and efficiently. The NVO will act as a coordinating and enabling entity to

foster the development of tools, protocols, and collaborations necessary to realize the full scientific potential of astronomical databases in the coming decade.

The NVO should not be viewed just as a new information infrastructure for data-rich astronomy. Rather, it will be *a comprehensive research environment for the new astronomy with massive data sets,* including data, tools, and services. The NVO goes beyond the existing structures in that it would provide new and increasingly needed technical and scientific functions, including unprecedented data fusion and data mining capabilities, instead of just the passive serving of limited data sets of the kind we have available now. It will make possible rapid querying of individual terabyte archives by thousands of researchers, enable visualization of multivariate patterns embedded in large catalog and image databases, enhance discovery of complex patterns or rare phenomena, encourage real time collaborations among multiple research groups, and allow large statistical studies that will for the first time permit confrontation between databases and sophisticated numerical simulations. It will also facilitate our understanding of many of the astrophysical processes that determine the evolution of the Universe. *It will enable new science, better science, and more cost effective science.*

The subsequent chapters and appendices of this report expand in some detail on the background, the scientific motivation and needs, the implied technical functionalities, and the possible implementation paths for the NVO.

1.4 Building on the Existing Foundations

This novel enterprise rests on a solid foundation of successful and scientifically highly productive applications of information technology in astronomy. It represents both a qualitative and quantitative leap ahead, but it is one based on an extensive existing information infrastructure and an active and vibrant community of developers and users of massive data sets and data services. The existing capabilities in astronomy, especially those developed over the past decade of so, are truly remarkable. The NVO will continue on this path, enhancing the state of the art and opening major new vistas for astronomy.

Appendix A gives a detailed account of the existing astronomical data centers and archives, together with ongoing projects and activities relevant for the development of the NVO within the U.S. and their counterparts worldwide. NVO-related developments already have developed a healthy momentum, but a vigorous, focused NVO effort is necessary in order to achieve the full scientific potential offered by the combination of information technology and data sets of increasing size and complexity.

1.5 Broader Significance and the Synergy With Other Disciplines

Astronomy is not alone in facing these problems. The technological challenges for the NVO are similar to those facing other branches of science, such as high energy physics, computational genomics, global climate studies, geophysics, oceanography, etc. The development of tools and techniques to handle astronomical datasets of this size will clearly have to call upon new developments in computer science and will have applications to fields outside astronomy. The full power of these databases cannot be tapped without the development of new tools and new institutional structures that can consolidate disparate databases and catalogs, enable access to them, and place analysis tools in the hands of a broad community of scientists. Research and

development of information systems technology is already underway in areas such as statistical analysis and data mining of large archives, distributed computational grids, data intensive grid computing (data grids), and management of structured digital information (digital libraries). Much of this research is relevant to the problems faced by the NVO. Information technology and data management throughout the sciences will both advance, and be advanced by, the NVO.

This creates a natural synergy between the NVO and applied computer science, statistics, and information technology. Strong partnerships between these disciplines, with the resulting benefits for all involved, are a natural way of developing and using the NVO. The computer science community already recognized this in a very tangible manner: much of the current funding for the development of the NVO comes from the NSF Information Technology Research program. Collaborative programs with statisticians are also under way.

The NVO concept also represents a powerful engine for the democratization of science. Making first-rate resources (the data and the tools for their exploration) available to the entire astronomical community via the Web would engage a much broader pool of talent, including scientists and students from small educational institutions in the U.S. and from countries without access to modern or large telescopes or space observatories. Human intelligence and creativity are distributed much more widely than the technological resources of cutting-edge astronomy (or indeed any science). Who knows what great ideas would emerge from this pool? This opening of a scientific opportunity to everyone with Web access may have a very significant impact in the long run.

2 Motivation: A New Scientific Paradigm

In this Chapter, we examine the scientific motivation for the Virtual Observatory. We achieve this through the discussion of an array of scientific scenarios, all of which are presently unachievable under the present mode of performing astrophysical research, but will be possible because of the Virtual Observatory. Such case studies demonstrate the necessity for a Virtual Observatory as well as the creation of a new research paradigm, which embraces the latest advances in information technologies, statistical knowledge discovery and distributed computing. We emphasize that in proposing the creation of the NVO, *we are not advocating any specific science; rather, the NVO will enable and facilitate certain kinds of science*. In summary, the NVO will provide advance our ability to investigate the Universe in three key ways:

- By facilitating a vast increase in the efficiency of doing science on massive data sources,

- By producing a truly coherent panchromatic view of all sources in the Universe, and

- By allowing for the seamless integration of observations and theory.

We discuss these three advances in greater detail below.

2.1 Efficiency of Performing Science on Massive Data Sources

The Virtual Observatory is designed to empower all users and increase their productivity by removing the inefficiencies associated with data management and large-scale analysis of present astronomical databases. This will allow the scientists to do what they do best: To ask the right scientific questions of the data and then interpret and publish the subsequent answers to their

queries. Therefore, the NVO astrophysicist will focus on understanding the Universe rather than hunting and gathering the data they need.

As an illustration of the expected increase in science productivity, we discuss the impact of the NVO on the search for distant clusters of galaxies in the X-ray satellite archives, which is vital for constraining our cosmological models. Table 1 shows the Past, Present and (possibly) Future state of this research area. Before the NVO (columns labeled Past and Present in Table 1), only hundreds of distant clusters have been found. This whole exercise has involved tens of researchers around the world who have dedicated decades to constructing these samples of clusters. The scientific return of this work has been somewhat disappointing since there remains significant ambiguity over the value of Omega (the matter density of the universe) derived from these cluster surveys. We await a high precision measurement of this fundamental parameter which will be possible through the NVO since researchers will be able to cross-correlate all available X-ray and optical survey data, and through the use of innovative statistical tools, extract all possible clusters from these growing archives. This utopia is displayed in the Future column of Table 1 – we do not need to collect any more data since there is a predicted 10000 clusters waiting to be discovered in the data we have today. The only hurdle in finding these clusters is the laborious nature of the optical follow-up observations, which will not be a problem for the NVO.

Table 1 The Efficiency of the NVO

	PAST (Pre-NVO)	PRESENT (Pre-NVO)	FUTURE (Post-NVO)	Projected improvement
No. of Clusters	100 (EMSS)	~300 for SHARC, RDCS, WARPS, CfA, MACS	~10000	100x
Data Archives	Einstein	Einstein, ROSAT	ROSAT, XMM, Chandra, ASCA	2x
Manpower & time to completion	5 people for 10 years	30 people for 5 years	1 grad student for six months	100x
Selection Function	Analytical calculation	Monte Carlo simulations	Simulated universes	?
Science Returns	Omega to 100% accuracy (0.1 to 1)	Omega to 50% accuracy (0.2 to 0.4)	Omega to 1% accuracy	20x

Beyond just expanding the size of distant X-ray cluster surveys, the NVO will change the way we test and understand these surveys. Today, the selection function of X-ray cluster surveys is derived via rudimentary Monte Carlo simulations i.e. by adding fake, idealized clusters to the

data and simply determining the success rate in detecting them. In the future, the NVO will make available simulated universes, constructed "on-the-fly", from numerical simulations and semi-analytical models of galaxy formation. Observers will simply download these simulated datasets with their real data and treat them like real observations, thus determining the sensitivity of their algorithms as a function of all the possible cosmological and observational parameters.

In the last column of Table 1, we attempt to quantify the level of improvement gained by the NVO compared to present methods of finding distant X-ray clusters. Overall, the NVO represents a many orders of magnitude improvement over present-day capabilities, which is an impressive return on the investment. Table 1 provides only one example of where the NVO will greatly improve science productivity: Similar improvements will be obtained by all researchers using the NVO as they search for quasars, stars and planets etc. in massive data sources.

2.2 Examples of NVO Science

Much of the power of the NVO will be in the capabilities of accessing, rapidly and efficiently, multiple data archives across a broad range of wavelengths and producing federated (or joined) data sets, which would enable new insights not achievable through any of the data sets taken individually. Here we offer some illustrative examples of this capability. Appendix B contains additional use cases.

2.2.1   Examples of Possible NVO Queries

To further illustrate the expected increase in scientific efficiency of the NVO, as well as the panchromatic and multidimensional nature of NVO investigations, we provide here a handful of possible science queries issued by future users of the NVO in pursuit of scientific projects like those discussed above and in Appendix B. Although these queries appear to be conceptually straightforward, they do require vast computational effort including the detection and federation of different datasets, automated statistical analyses and novel visualization of the final products. It is worth stressing that the results of these queries will still require significant interpretation by the user; therefore, the NVO is not replacing the scientists, but making them much more efficient. The NVO will release them from the burden of data management and manipulation to spend their most precious resource, i.e. their time, on the physical meaning of their results.

1. To quantify the star-formation rate of galaxies, give me all galaxies that have a detected H_a line with equivalent width greater than 5 Angstroms. What fraction of these galaxies is detected in the radio?

2. To identify possible AGNs, use the emission line ratios measured from SDSS and 2dF galaxy spectra to find low-luminous AGN candidates. Return their colors and emission line parameters.

3. To automate the identification of X-ray sources, return the distribution of the optical-to-X-ray fluxes for DPOSS sources detected within the error circle of XMM serendipitous sources.

4. To find gravitational lenses, identify all elongated sources within 30 arcseconds of the brightest galaxies in known clusters of galaxies. Return their colors, positions and atlas images to the user.

5. To discover Brown Dwarfs, select all stellar objects for which 2MASS JHK colors are like those of an A star and there exist overlapping optical catalogs such that V-J > some limit. Include objects for which an optical survey has been conducted but no object was detected at a specified sensitivity limit. Return atlas images in all filters of such objects, including optical non-detections.

6. To find variable objects, overlay this image from HST with maps from the VLA. Show all known sources in this area with links to data and references. Do any of the objects have multiple epoch observations?

7. To quantify the structure of our Galaxy, compute the correlation function of all known A stars correcting for edge effects and incompleteness.

8. To constrain cosmological models, return 100 simulated surveys of AGNs and galaxies over the SDSS and VIRMOS areas of the sky. Use a user-defined biasing prescription to place galaxies and AGNs in the dark matter haloes. Impose a surface brightness limit of 28 magnitudes per square arcsecond and return RA, DEC and multi-color magnitudes for all sources.

9. To plan for a new satellite mission, find all quasars that possess both infrared and UV spectra with resolutions of greater than R=1000. Return the number and spectra.

10. To study solar active region velocity fields, select times and regions between the latitudes of +15° and +30° North where the average solar magnetic field strength is greater than 150 Gauss. Return a time series of spectra obtained around the Fe I 5576 line at these locations.

2.2.2 A Panchromatic Universe: Explosion in Science

For the first time, the NVO will open up a panchromatic and multidimensional view of the universe for all possible sources to all possible users. This will cause an explosion in scientific discoveries and, to illustrate this breadth of possible NVO science, we outline in Table 2 an array of scientific scenarios that will greatly benefit from the creation of a NVO. We also present below one of the examples in Table 2 but discussed in much greater detail to illustrate the great

Table 2 The Breadth of Science with NVO

Planetary/Solar System Science	Stellar & Galactic Astronomy	Extragalactic Astrophysics	High Precision Cosmology
Understanding the Solar Wind	An observational model for our Galaxy	Testing the evolution and unification of AGNs	Clusters as cosmological probes
Asteroids, Comets, and Kuiper Belt Objects	Tracing the Galactic halo using stars	Galaxies in N-imensional Parameter Space	Gamma-Ray Bursters and Cosmic Star-formation
Search for Extra-Solar Planets	Massive Catalogs of Peculiar stars	The large-scale structure in the Universe and Cosmic Web	Cosmic Microwave Background and Foreground contamination

potential impact of the NVO, while in Appendix B, we discuss many more of the scenarios in Table 2 in greater detail. These scenarios will demonstrate the power of the NVO in stimulating new science.

Appendix B gives a number of examples case studies corresponding to the entries in this table. Here we illustrate one of these cases.

2.2.3   A Case Study: Testing the Evolution and Unification of AGNs

Active galactic nuclei (AGN) are the most energetic single objects in the Universe, with total energy production in excess of 10^{60} ergs. These objects produce copious amounts of non-thermal emission, magnetic fields, and relativistic particles, and in addition they often are the source of highly collimated relativistic outflows that persist for 100 million years or more. The energy production rate for these objects can be equal to 100 times the total radiated energy of a typical galaxy, yet all of this energy is produced in a volume much less than a cubic parsec. The astrophysical processes at work in AGNs are thus some of the most mysterious in the Universe, and a complete understanding of how they function will lead to new insights into the behavior of black holes, the evolution of galaxies, the general theory of relativity, and perhaps new physics.

Thus a complete understanding of the nature and characteristics of the different types of AGNs we observe in the Universe is a fundamental goal in astrophysics. Yet our understanding of these objects is severely hampered by the large dynamic range witnessed in their properties, e.g. they exhibit a wide range of redshifts, variability, obscuration and spectral shapes, thus giving rise is an array of different colors across wavelengths from x-ray to radio. Astrophysicists are attempting to construct larger, more complete, samples of AGNs in order to test many of the physical hypotheses about AGNs, including: a) Determining if the observed properties of different types of AGNs are dominated by factors like orientation, environment and/or obscuration, thus testing the validity of the so-called "unification model"; b) Understanding the evolution of the AGN luminosity function, and, in particular, separating number evolution from luminosity evolution; c) Constructing the AGN luminosity function at different wavelengths to understand the global evolution of AGN properties.

Outline of the Project:

- Federate a large number (N) of surveys covering the same significant area on the sky, and spanning a large wavelength range (X-ray through radio).

- Include meta-data information (e.g. flux limit, areal coverage, passbands) so that the survey selection function can be accounted for in subsequent analyses.

- Identify AGN candidates in this joint dataset. Note that although confirmation might require new observations, the process could be carried out in a statistical sense, resulting in a probability that any particular object is an AGN. One can further classify the object in particular, well-established, AGN classes.

- Compute derived quantities for each AGN e.g. orientation, obscuration, environment etc.

- To test the "unification model" of AGNs, identify statistically significant correlations between objects in an N-dimensional parameter space i.e. look for joint correlations between classes of AGNs and their orientation, obscuration, environment and spectral characteristics.

- To compute the AGN luminosity function, determine the survey completeness and volume given the joint selection function of the different surveys. Perform a statistical test to determine the degree of evolution as a function of several parameters including redshift, environment etc.

- Construct simulated catalogs of AGNs based on cosmological simulations and semi-analytical theoretical models of AGN formation and biasing. Such simulations should contain the same selection functions as the real data and, when compared to the data, be used to constrain the input models and cosmology.

NVO Functionality Required:

- Federation of relevant surveys including cross-identification of objects in multi-wavelength surveys and interchange/merging of meta-data into a standard format.

- Statistically robust clustering algorithms to identify classes of objects in multi-dimensional data sets. This will include both supervised analyses (in which astronomical knowledge guides the definition and analysis) and unsupervised analysis (in which new patterns are recognized).

- Automated search for statistically significant correlations within the datasets and visualization of the results to facilitate user feedback, change search parameters, etc.

- Ability to allow user-defined functions and algorithms to be run on remote datasets e.g. to compute the local environment of AGNs, to determine the luminosity function of AGNs.

- Creation of simulations on-the-fly, based on user supplied inputs or algorithms. Application of observation selection functions to these idealized simulations to generate simulated catalogs with the same quantities as the real data. Allow user-defined biasing prescriptions for the placement of AGNs in the dark matter haloes.

2.3 The Marriage of Theory and Observation

The technological advances that have enabled the NVO have also provided the means for a truly revolutionary blending of theory and observation that has been, until now, impossible. For the first time in the history of astronomy, it is possible to carry out detailed calculations in theoretical astrophysics that address problems whose solutions have eluded all of astronomy. The exponential growth in computational capability has enabled this remarkable achievement, particularly through the establishment of national supercomputing centers and through the development of distributed parallel computing facilities. Very rapid progress is being made in studying nonlinear phenomena such as supernova explosions, the formation and development of accretion disks, and the evolution of relativistic outflows from active galactic nuclei. The ability to calculate enormously complex problems has also led to revolutionary progress in studying N body systems, including following the evolution and interaction of every star in a globular cluster, and to the development of very complex stellar evolution codes that can reproduce the spectral energy distribution of entire galaxies as they evolve.

Perhaps one of the most exciting syntheses of these remarkable advances in computational astrophysics is the ability to follow the evolution of the entire universe from the original primeval density fluctuations through the formation and evolution of galaxies and clusters of

galaxies. These elegant and intricate calculations involve a symbiotic combination of general relativity, cosmology, nonlinear gravitation, N body interactions, gas dynamics, and stellar evolution. They include the presence of dark matter and "dark energy" as well as ordinary baryonic matter. With current technology, theoretical astrophysics is coming close to being able to model the Universe from the first few seconds after the Big Bang until the present day.

All of these investigations, from star formation through black hole dynamics to the evolution of the Universe, produce massive data sets. Moreover, the basic physical quantities contained in these data sets can be converted into directly observable quantities: magnitudes, metallicities, line strengths, velocities, etc. Thus for the first time it will be possible to make direct comparisons between complex theoretical calculations and large, statistically significant observational databases. Both observational and theoretical data sets will be readily accessible through the NVO structure along with the tools necessary to compare theory and observation. Thus the NVO will make possible for the first time an interplay between theoretical models and observational data on a truly meaningful scale; observers will be able to request simulated universes based on a suite of input parameters and then download the simulated datasets like real survey data. The NVO will make these numerical simulations available to all users, which will result in the rapid refinement and verification of theoretical modeling, which, in turn, will lead to rapid advances in our understanding of complex astrophysical processes.

Theoretical astrophysics will also make a major contribution to the educational and public outreach efforts that will be coordinated through the NVO. The complex numerical simulations outlined above produce many striking and insightful images, as well as graphic animations that enable the viewer to grasp the physics at work during the calculation. Such images and animations will provide an extremely valuable resource for both formal educational applications and for general public interest.

2.4 Prospects for a Qualitatively New Astronomy

In this Section, we expand our scientific discussion to emphasize the ability of the NVO to enable a qualitatively new form of astronomical research. This is based on the concept of opening up the whole area of observable parameter space to all astronomers. We discuss this in greater detail here.

Every astronomical observation and survey covers a portion of the observable parameter space, whose axes include the area coverage, wavelength coverage, limiting flux, etc., and with a limited resolution in angular scales, wavelength, temporal baseline, etc. Each one represents a partial projection of the observable universe, limited by the observational or survey parameters (e.g., pixel sampling, beam size, filters, etc.). Every astronomical data set samples only a small portion of this grand observable parameter space, usually covering only some of the axes and only with a limited dynamical range along each axis. Every survey is also subject to its own selection and measurement limits. Surveys thus represent hyper-volumes in the observable parameter space. Individual sources represent data points (or vectors) in this multi-dimensional parameter space. So far we have sampled well only a relatively limited set of sub-volumes of this observable parameter space, with a much better coverage along some of the axes than others. Some limits are simply technological or practical, but some are physical, e.g., the quantum noise limits, the opacity of the Earth's atmosphere, or the Galactic interstellar medium.

Federating multiple surveys that sample different portions of the observable parameter space can provide a much more complex and complete view of the physical universe. The simplest and most traditional, yet very powerful, manifestation of this process is cross-identification of observations at different wavelengths.

A parameter space representation of the available data shows us not only what is known about the universe, but also what are the unexplored areas where genuine new discoveries may be made, and also what is knowable given the technological or physical limits of our measurements. Now for the first time we have the adequate data and technology for such a global, empirical approach to the exploration of the universe.

The unexplored regions of the parameter space are our *Terra Incognita*, which we should explore systematically and where we have our best chance to uncover some previously unknown types of objects or astrophysical phenomena – as well as reach a better understanding of the already known ones. This is an ambitious, long-term program, but we believe that with the data sets already in hand it is possible to make some significant advances.

2.4.1 Historical Analogs

Historical examples abound. The discovery of quasars resulted when some of the first well-localized radio sources were identified in the visible light with what appeared as otherwise non-descript bluish stars – yet they represented a new, spectacular natural phenomenon, and the science of astronomy was changed fundamentally. A similar revelation happened when the sky was mapped for the first time in the far-infrared by the IRAS satellite: it was discovered that the most luminous objects in the nearby universe, identified optically with some irregular galaxies previously considered as mere curiosities, have their powerful energy sources hidden behind the veils of opaque dust. It is now believed that at least half of all star-formation regions and a large fraction of all AGN in the universe are escaping detection in the standard visible-light techniques due to such obscuration. Cosmic gamma-ray bursts (GRBs) are another spectacular example, where the breakthrough in the understanding of a new astrophysical phenomenon came from insights gained at other wavelengths, including x-ray, optical, and radio: a 30-year cosmic mystery was resolved by combining the data from a range of wavelengths. Essentially all of the x-ray astronomy illuminated the universe in a new light: clusters of galaxies are primarily x-ray emitting objects, and we have gained great insights into a range of variable stars and AGN from the x-ray point of view, with information not available in any other wavelength regime. Even the well-known supernova explosions emit about 99% of their energy in the form of hard-to-catch neutrinos; the spectacular fireworks we see represent only about 1% of their total energy budget.

These remarkable discoveries represent simple examples of what is possible when data from a range of wavelengths are combined in a systematic way. *Fundamental new phenomena are found, and new insights are gained in the types of objects and processes already known to exist.* We can surely expect even more spectacular discoveries and insights as we start to combine and explore in detail the new generation of massive digital sky surveys, leading us to a better, more complete and deeper understanding of the physical universe.

2.4.2 The Time Domain

A relatively poorly explored domain of the observable parameter space is the variability on the sky, especially at faint flux levels, at every wavelength. While a number of variable types of

objects are already known, including many types of variable stars, quasars, novae and supernovae, GRBs, etc., we know very little about the time-variable universe in a systematic way. There are already some puzzling phenomena found, e.g., the fast, faint optical transients, which may or may not be associated with distant supernovae, and the mega-flares on otherwise apparently normal stars, which brighten by a factor of a few hundred for a period of hours or days, for as yet unknown reasons. Since the time domain has historically been vital for solar physics research, the incorporation of solar experience could bootstrap this aspect of the NVO toolbox.

2.4.3 The Low Surface Brightness Universe

The most difficult problems to overcome in astronomy are bias and selection effects. What we can observe literally becomes a task of what we can find in our catalogs and surveys. Bias is well known in extragalactic research; bias to apparent brightness, bias to distance, bias to optical emitters. However, less known is the cumulative effect that size, luminosity and density have on our ability to understand the full range of galaxies in our Universe. One of the primary goals of the NVO is to enable new discoveries and new science through the use of novel methods of data analysis and interpretation. Most data mining projects focus on extreme objects (the reddest or most luminous) due to the nature of isolating a sample of objects with a range of characteristics. But, the NVO allows for the reverse to take place, the mining of parameter space, in particular the exploration into regions where bias and selection effects have prevented analysis. The NVO allows researchers to combine datasets with new algorithms as they are developed with emphasis on unusual domains of parameter space. One such domain is the region of parameter inhabited by objects that have low contrast to the background. This region of parameter space has been poorly explored to due the computational difficulty in the detection of these diffuse, low surface brightness galaxies. By joining tools and datasets, NVO researchers will have unprecedented ability to explore the full parameter space of galaxy size, luminosity and surface brightness, at various wavelengths, to quantitatively, and systematically, determine the physical characteristics of galaxies. This process not only leads to direct discovery and new knowledge, but also serves to guide future studies and improve the efficiency of planned observing programs.

2.4.4 Searches for New Types of Objects and Phenomena

Probably one of the most interesting new kinds of science enabled by the large data sets is the search for rare and new types of objects or phenomena, discovered as outliers in some parameter space of observed. This is an "organized serendipity" approach to exploration of large data sets and a prime example of data mining applications in astronomy.

A typical NVO data set may be a catalog of $\sim 10^8$-10^9 sources with $\sim 10^2$ measured attributes each, i.e., a set of $\sim 10^9$ data vectors in a ~ 100-dimensional parameter space. Objects of the same physical type would form clusters or sequences in this parameter space. The most populous clusters will be those containing common types of objects, e.g., normal stars or galaxies. But the unusual and rare types (including possible new classes) would form sparse clusters or even be just individual outliers, statistically distinct from the more common ones. Rare objects may be indistinguishable from the more common varieties in some observable parameters, but be separable in other observable axes. This approach has been proven to work and is now used on an industrial scale to find high-redshift quasars and brown dwarfs in SDSS, 2MASS, or DPOSS.

We are mainly interested in a search for previously unknown types of objects or rare phenomena associated with known types, manifesting themselves through "anomalous" properties in some large parameter space. If some type of an interesting object is, for example, one in a million or a billion down to some flux limit, then we need a sample of sources numbering in many millions or billions in order to discover a reasonable sample (or even just individual instances) of such rare species. Obviously, we don't know what will be found — but that is what makes this study exciting. This is exactly the kind of new science enabled by massive data sets.

2.5 Benefits for Other Sciences

2.5.1   Applied Computer Science

The NVO provides a wonderful source of computer science challenges, and a laboratory for testing new techniques. The Virtual Observatory is a perfect application of Grid computing technologies. Federating the world's astronomy archives is a good application of the metadata, web service, and distributed query technologies computer scientists have been working on for decades - but these challenges are at or beyond the limits of what we can do today. Managing the Petabytes of NVO data, and performing cross-correlations among objects will require new approaches and algorithms. Some examples are:

- Describe the NVO contents so that scientists can discover the information they need. This will probably be some form of XML schema that maps to each member of the federation, and also a directory service that catalogs the NVO holdings;

- Provide location-transparent query facilities that produce an answer set from data pulled from various archives. Make this query facility fast enough so that it can be used interactively;

- Visualize answer sets so that scientists can interactively explore the datasets;

- Provide near linear-time approximate algorithms to do cross-correlations among these huge datasets (current algorithms are polynomial and so fail on large datasets);

- Develop algorithms for automated discovery in high dimensional data.

Solutions to these challenges will likely benefit all the sciences, but the NVO provides an excellent testbed to focus these efforts.

2.5.2   Applied Statistics

Large-scale astronomical datasets present some important challenges to the community of mathematical statisticians and applied mathematicians. While multivariate analysis is a well-developed field, there are few methods that treat two conditions that will commonly appear in NVO datasets: known heteroscedastic (i.e. different for each object) measurement errors and censored data (i.e. upper limits due to non-detections) in many variables. Many specific NVO problems will engender real statistical challenges: characterizing the anisotropic nonlinear clustering of galaxies; classification of large collections of optical spectra; parameter estimation for nonlinear astrophysical models fitted to heterogeneous datasets; and so forth. Close collaboration between NVO astronomers and statisticians will be needed to fully address these

complex problems. The statistical community in turn may well be enriched by the challenge of these astronomical problems. Statistical analysis of billions of objects with complex and heterogeneous attributes, presents entirely new challenges for both statisticians and computer scientists.

2.5.3   Interdisciplinary Exchanges With Other Data-Intensive Fields

Biology is experiencing a similar data-flood as astronomy and astrophysics. Their data storage and analysis problems are very similar to the problems we face in astrophysics and thus we can hope for much cross-disciplinary research. For example, the detection and characterization of proteins as imaged by digital microscopes is analogous to the detection of galaxies as seen by a CCD detector. Recently, at Carnegie Mellon University, biologists have successfully begun to use SExtractor, a software package designed for astrophysics, for their analysis of protein images. This illustrates the universal nature of the Virtual Observatory since it could be used to store, manipulate and analyze any digital data.

3   Requirements and Capabilities of the NVO

The NVO will take on a variety of roles and responsibilities that build upon the extant data archives and information services:

- Enabling new science from large, highly distributed datasets
- Facilitating new surveys and missions
- Fostering NVO-enabled research

Science requirements in particular are key drivers for the definition of technical requirements. This will also automatically satisfy the second role. The NVO will not be developed in a vacuum, however, and will be able to utilize and co-opt existing and the expected future technological advances made in the IT arena by both academia and industrial partners.

3.1 Enabling New Science

In the previous chapter and in Appendix B we discuss a number of scientific projects that are facilitated by the NVO. These science projects require a suite of functional capabilities that the NVO must provide in order to be scientifically useful:

- Provide capabilities for discovering what data are available to the NVO user, and for easily incorporating new data into the NVO framework.
- Provide seamless access to globally distributed data, whatever its type (observational or simulated, images, spectra, catalogs, etc.).
- Provide mechanisms for federating globally distributed data, whatever its type.
- Provide seamless access to any online service, wherever it is located.
- Develop universal standards for archiving future data sets.
- Develop analysis toolkits that can be used as is, or extended as needed, which facilitate processing of large datasets, including catalogs, images, and simulated data.

- Develop new techniques for visualizing large quantities of data, including catalogs, images, or simulated data.

- Link with existing and future digital libraries and journals.

These capabilities, which might not seem revolutionary on their own, are qualitatively different from what is currently available due to the size, dimensionality, and complexity of the current and forthcoming datasets that will populate the NVO. The NVO will progress along an evolutionary path, slowly adding new functionality, and eventually all of these capabilities will become possible. In the following subsections, we explore the design implications of these requirements. In Appendix C, more detailed design considerations are presented.

3.1.1   Resource Discovery and Data Publication

The NVO will be a dynamic environment in which new archives appear, existing archives grow and evolve with time, and old archives disappear. The NVO must provide a dynamic mechanism for discovering which archives exist and what types of data they contain. Each survey represents a hyper-volume in the multidimensional parameter space of area coverage, wavelength, resolution, time, etc. This information can be an integral part in the formulation of any scientific query and so needs to be made available as part of the resource discovery mechanism.

The resources of the NVO can be cast in the form of a large directory, with individual archives or other resources being nodes within the directory. Because only a small number of these resources will be relevant to a typical scientific query, the directory must provide the ability to rapidly identify the relevant resources and ignore the rest. The boundary of the hyper-volume that a survey occupies within a multidimensional parameter space is seldom describable in simple terms; thus, methods must be developed for approximating the boundaries in a concise fashion. For example, sky coverage can be described using a hierarchical triangulated mesh scheme.

Publication of new data to the NVO must be straightforward. However, new datasets must be conformant to a set of formatting standards that should be verified. The new datasets must provide at least a minimum description of the essential properties of their contents-including information about the quality and completeness of the data-with sensible default values used for unspecified properties.

The NVO must be able to discover not only data archives but also computational resources. For example, some operations may require extensive computations that require replicating data to other sites. For certain datasets (e.g., those hosted at sites with limited computing capability), it will be cost-effective to maintain replicas of those datasets, which the NVO must be capable of storing and tracking. Such replicas, however, must be tagged and versioned, since the original datasets may evolve in time. NVO must be able to discover where computational resources exist, what their capabilities are (storage, CPU, computing environment), and what bandwidth capability exists between them and the data archive. Problems requiring these capabilities are being addressed by a variety of grid computing projects (e.g., Griphyn), and NVO should be able to benefit from this experience.

3.1.2 Seamless Access to Globally Distributed Data

NVO data sets will not be archived in one physical location, but rather will be distributed at many facilities and managed and updated by domain experts. In order for scientists and computer programs to make use of such information, the data must conform to agreed upon metadata standards. Metadata describes the nature and structure of the data, such as the header records (keywords and values) in a FITS file. The FITS format, however, only defines the meanings (semantics) of data files for a few keywords (metadata concepts), focusing instead on the structure (syntax) of the information. A major challenge for the NVO is to broaden the semantic standards for astronomical data so that archives and catalogs can be queried using common terms, and query results can be interpreted by both machine and human readers without ambiguity. These standards must also encompass the results of theoretical simulations so that models and data can easily be compared. Initially the NVO metadata standards will be used to define interface layers that translate between the standard concepts and terms and site- or service-specific terms. Eventually the NVO metadata standards are likely to be adopted for direct use in image headers and table column headings, minimizing the need for interfaces or mediation programs. These standards should be created within the global VO context.

3.1.3 Federating Globally Distributed Data

Data federation is the ability to treat multiple, independent catalogs, databases, or data sets as an integrated whole. Combining existing datasets can create new knowledge, knowledge that does not require a telescope or a rocket launch. A prerequisite for data federation is interoperable metadata standards, but for large amounts of data, there are additional requirements. Caching and replication services can save the results of complex joins of multiple databases for later use. An efficient proximity join of a billion sources requires that the data be clustered so that nearby objects in the sky are also nearby in the data stream. The development and utilization of efficient indexing algorithms are therefore required.

3.1.4 Seamless Access to Services

The NVO will not only need to deliver data to end users, but also to send data to appropriate applications software that may or may not reside on the end users' computers. A data request could involve server-side computations (image registration, image subsetting, image mosaicing, source extraction, statistical tests) that are hosted on designated servers in the NVO grid. These are the functional equivalent of services like today's airline reservation systems-a user of Travelocity or Expedia does not know, or need to know, what systems are actually searching for flight availability or making reservations. The difference, however, is that the NVO will require moving potentially large amounts of data to the computational services. The matching of data to services will be based on the emerging Grid technologies.

3.1.5 Developing Universal Archiving Standards

The use of databases to organize, track and access astronomical data has grown significantly over the last 10 years. The growth has been driven by an increasingly largeNVOlume of data, the need to perform increasingly complex queries, and access to relatively low-cost commercial Relational Data Base Management Systems (RDBMS). The data holdings in existing archives are generally organized with a set of associated metadata sufficient to support queries about the

data holdings and to allow data set retrieval. Some services such as NED and SIMBAD provide query and retrieval access across selected multiple archives. These services are examples of the first steps towards a database architecture that supports general cross-archive and cross-catalog query and retrieval.

The NVO will begin by building on existing databases, archives, and catalogs to provide access to current space and ground-based data. This implies interfacing with a variety of underlying commercial and non-commercial database systems, metadata formats and structures, and suggests a query protocol sufficient to provide meaningful and sophisticated queries across a wide range of data sets. A set of minimum standards will be needed as part of the implementation. The standards will cover the data organization, metadata format and content, query protocol interface, and data delivery protocol. They will lead to an architecture that includes a translation layer between the NVO query protocol and the underlying databases associated with the existing archives. A layered approach is desirable since it minimizes change at existing provider's sites. The development of such standards is a critical function of the NVO because it will allow future data sets to be designed and generated as NVO-compliant, in-turn resulting in increased functionality for minimum cost. The standards will also enable the development of tools to allow increasingly efficient development of future NVO-compliant data sets.

The layered, standards driven database architecture described here, is consistent with an overall NVO architecture consisting of tools layered on a set of underlying services interfacing through a set of standard exchange protocols. It is likely-and beneficial-that future archive and database designers can make direct use of many of the NVO specifications, thereby minimizing development efforts and making interface layers extremely simple.

3.1.6   Developing Analysis Toolkits

Both the developers and users of the NVO will see the need for specific, but NVO-aware, data analysis tools. These might be specially tailored versions of more generic tools (i.e., adapting a general classifier to work on peculiar stars), or customized source extraction codes that are deployed to reanalyze large amounts of survey data. The community-based research support program is expected to foster the development of such toolkits, motivated by individual research projects but made available within the NVO framework for reuse and adaptation by other scientists.

3.1.7   Developing Visualization Capabilities

The vast quantities of information that the NVO will make available to astronomers around the world threaten to overwhelm the general user. In order to make sense of all of this information, new methods for visualizing both catalog data and large image mosaics on the desktop need to be developed and provided as part of the NVO infrastructure. NVO users will frequently be dealing with N-dimensional parameter spaces and will want to visually understand the correlation of parameters, and perhaps more importantly, quickly detect outliers from "normal" behavior.

3.1.8 Incorporating Digital Journals and Libraries

The scope of the NVO extends from the data archives and associated object catalogs to what has been written about these data and catalogs in the literature. We have already seen the value of cross-referencing from services such as ADS to the on-line astronomical archives, and vice versa, so that bibliographic and data searches can quickly locate the underlying information. Moreover, the published literature includes many small and mid-size tabular data sets, and many of these are already being incorporated into the on-line services of the ADC and ADS. There should be many entry points into the NVO framework for researchers: from published papers, back to the original data and supporting or complementary data sets; from data and catalog-oriented data requests, including references to the appropriate literature citations; and from outreach and education sites.

3.2 Technical Foundations

The fundamental technical challenges that must be overcome in designing and implementing the NVO can be summarized in three key areas: storing data, transmitting data, and processing data. Fortunately, the same fundamental driving forces of information overload that are leading our community to develop virtual observatories are also affecting other academic disciplines as well as the corporate world.

As a result, we are not forced to work in isolation, and we can and will adopt both technology and partners as are warranted. The primary current initiative that we foresee adopting is the concept of collaborative computing, which encompasses everything from peer-to-peer computing (which includes projects as diverse as SETI@Home, Napster, Gnutella, or Jabbar) to computational grids.

3.2.1 Computational Grids

In the IT world, Grid based computing is the latest buzzword. Behind the façade, however, Grid computing actually holds significant promise for providing an affordable, high performance, distributed computational infrastructure. Grid Technology allows a user or application (such as the NVO) to utilize highly distributed computational processing resources, data storage resources, and networking resources as a single system. One of the key concepts behind grid applications is "Virtual Data". This allows an application to interact with the Grid without actually knowing where the data resides, or is being processed.

Several projects are leading the way in developing Grid Technology, including Globus (www.globus.org), Griphyn (www.griphyn.org), and the International Virtual Data Grid Laboratory (www.ivdgl.org). Another prime candidate for collaboration is Teragrid (www.teragrid.org). Teragrid is a collaboration between NCSA, SDSC, Argonne National Laboratory, and Caltech to build a prototype Grid that includes Teraflop processing with Terabyte storage all connected by very high speed connectivity. Eventually, Teragrid is expected to expand to additional sites, providing even larger processing power and storage capacity. There is an expression of interest from these projects to collaborate with the NVO.

3.3 Facilitating New Missions and Surveys

Currently all new NASA Space Science missions are scoped to include some level of data archive, or as a minimum the provision of a mission archive to NASA at the conclusion of the mission. This is clearly a critical requirement given the large investment in space science missions, and leads to long-term accessibility of data. Efforts are made to ensure conformity with common standards, e.g., FITS, and data are required to be passed to long-term archive sites (e.g., the HEASARC for high energy data) once the mission is complete. Providing an NVO infrastructure that missions can build to, as part of their development, will reduce costs by providing templates for the generation of standards data products. These include the data, catalogs, calibration transformations, data base designs, and a library of existing analysis and display software. The NVO will also reduce the cost of accessing future data sets by providing standards for electronic publishing of data sets.

An important component of the NVO will be software simulations and models. These models will be both generated on the fly, and made accessible in pre-calculated forms. Access to such models will reduce the cost of planning new instruments and missions. One example is the provision of tools for the generation of star catalogs that meet specific magnitude, color, and astrometric criteria. Simulations of instruments for various mirror configurations are another important aspect of the development phase of new programs. The NVO will play a role by providing access to models and simulations including ray traces, astronomical simulations for point and extended sources, a variety of spectral or time domain parameters.

Future datasets will be more cost-effective to generate and access through the NVO because they can be inherently created as NVO-compliant. The compliance will be possible through the development of a series of standards as discussed in Section 3.1.

3.4 Fostering NVO-Enabled Research

Major NASA missions such as HST and Chandra have high scientific productivity not only owing to the excellence of the observing facility, but also due to the strong support of the scientists who use the telescopes through their associated research grants. Similarly, the productivity of the NVO will benefit greatly from a dedicated grants program for its scientific users. Funding for NVO science programs could come in several forms.

Tool Development. There should be regular "AOs" for opportunities to build software tools that utilize the NVO infrastructure. They would be delivered to the NVO for wider use by the community and would follow standards defined by the NVO. It is important that these tool-building opportunities cover a wide range of possibilities and engage a large part of the community. A strong science enabling case for each software tool must be made, but they will be general user facilities that the entire community can use to do research.

Focused Research Programs. There should also be regular AOs that provide funding to use the NVO for specific research projects with a well-defined goal. These programs might have a software component as well, but need not, and would be selected on the basis of scientific excellence.

Fellowship Programs. Again following the successful pattern of the Hubble Fellows Program, NVO will benefit from a funded fellowship program that engages young scientists in

the use of NVO facilities. This inherently interdisciplinary program should also include fellowships (or structured programs) for undergraduates and graduate students, providing early exposure to the relevance of information technology research generally, and NVO-enabled science specifically.

4 Education and Public Outreach

NVO will establish an effective EPO program that builds upon existing efforts, taking advantage of the unique aspects of the NVO data and framework, as well as new information and communication technologies that will be developing rapidly over the next decade. To be most effective, the EPO component of NVO needs to be built in from the very beginning of the project design.

4.1 A need for information technology and science literacy

Knowledge of science, mathematics and technology provides the competitive edge in today's society. Science and technology can act as key differentiators in our world's rapidly changing economy. The K-12 and general public communities need long-term, efficient access to appropriate science and technology resources; these must be integrated into school curricula, programming of informal science education centers, the media, etc.

In concert with the integrative nature of a virtual observatory that seamlessly cuts across distributed data archives, the NVO EPO effort will be designed to provide the user with an integrated view of the Universe - a value-added system that goes beyond traditional on-line portals to the various datasets. 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.

Additionally, NVO provides an opportunity to expand the existing NASA and NSF education and outreach efforts by identifying new sectors of society with a long-term interest in science and technology literacy. The OSS EPO program, for example, has primarily focused on the core space science areas. The NASA NVO effort provides a unique opportunity to enhance technology literacy in a broad sense, over and above basic interest in astronomy and space science. The NVO EPO program will identify computer science and technology innovations and practices that can be integrated into education outreach programs. For maximum impact, this will require close NSF-NASA education and outreach coordination, as well as adherence to National Education Standards for science, mathematics, and technology (National Science Education Standards, Project 2061 AAAS Benchmarks, National Council of Teachers of Mathematics Education Standards; International Technology Education Association Standards). This approach will facilitate integration of NVO resources into the pre-college curriculum.

4.2 Components of a Successful Program

4.2.1 Identification of NVO users

The success of the NVO EPO program will be dependent on identifying potential users and their needs. NASA and NSF EPO efforts have traditionally served pre-college and college educators and students as well as informal education institutions such as museums and

planetaria. There are a number of additional potential user communities that reach beyond the traditional audiences served by NASA and NSF data-driven initiatives. These communities include amateur astronomers, educators in teacher preparation programs, the art and entertainment communities, and libraries. Assessing the particular needs of the entire NVO customer base will be critically important to ensure that NVO resources and data access on-ramps are responsive to the varied requirements of different groups. This strategy will maximize the likelihood that NVO resources will actually be used and will be of benefit to the target audiences.

Digital images of interest to the K-12 community and general public are generated by a variety of science disciplines other than astronomy, as well as the arts and entertainment. There is already a great deal of interest about space science imagery from communities as diverse as art museums, public libraries, major Internet portals, and the media industry. Many NASA images are acquiring iconic status within the advertising and media industry. NVO will be able to reach beyond existing EPO efforts by allowing user communities to build unique knowledge of the science behind the images - the kind knowledge that only NVO will make possible through comparison and analysis of data from various archives.

4.2.2   Example Partnerships

The following examples illustrate that NVO EPO could offer innovative possibilities for new partnerships that could be integrated with existing efforts.

The Arts: This example would involve NVO providing imagery and tools through existing distribution systems that link art museums and art schools internationally. Foundations such as the Getty Foundation, and corporations such as Corvus, have made long-term commitments for the infrastructure and content of imaging data within the arts. A number of Internet 2 projects have been established in this area, with investments within the art world from companies such as Intel and Microsoft. Particularly, at the lower grades it may be possible to articulate an integrated approach that allows NASA imagery to be integrated into the needs of art curricula.

SETI@home: The SETI@home project, offers some obvious opportunities for innovative ways to disseminate NVO EPO materials. The SETI@home website currently receives 120,000 separate visitors daily, and SETI@home is used by 7000 school groups involving some 100,000 school children. SETI@home has captured the public imagination internationally and the number of users is steadily growing. Each set of data processed by a person who has installed SETI@home on their home computer is linked to a specific position in the sky. NVO EPO could routinely provide imagery of the portion of the sky being processed by that individual's home computer, along with associated educational materials. A recent survey of K-12 educators who use SETI@home revealed that the most wanted resources by this population are digital imagery of sky objects and space science data for use in the classroom to engage students in the process of science.

Pre-service Teachers: Another example involves using NVO to enrich and influence the training of future teachers of science, mathematics, technology, and the arts/humanities. Well-prepared teachers will be equipped to integrate technology into instruction and impact student learning, as they become practicing teachers in the K-12 community. In addition, technology-savvy teachers can serve as mentors to their peers and act as agents of change for enhancing the effective use of technology in the classroom. Enhancing teacher preparation, as suggested here,

in part responds to the recommendation by the aforementioned NAS Decadal Survey to develop partnerships between Astronomy Departments and Schools of Education for the purpose of improving science courses taken by future teachers.

Amateur Astronomers: The amateur astronomy community in the United States represents a valuable, largely untapped source of expertise, energy and enthusiasm for engaging the public, communicating astronomy and space science research, and conducting public outreach activities at many different levels. It is estimated that there are between 300,000 and 500,000 amateur astronomers, including about 20,000 - 30,000 "affiliated" amateurs (defined as those belonging to one of more than 250 astronomy clubs in the U.S). NVO will allow the amateur astronomy community to participate in the process of research and discovery through access to data and analysis tools. NVO will immensely enhance the knowledge of amateur astronomers, increasing their support and contributions to astronomy as well as dramatically increasing amateur-professional collaborations. Amateurs also interact with the K-12 community and community-based organizations such as boys and girls clubs, libraries, etc., and facilitate access to, and understanding of, space science and astronomy content, data, research discoveries, and appropriate materials. NVO will empower the amateur astronomy community to share the process of science with broad audiences, serving as a bridge between the professional community, educators, and the public.

Students: This report recommends the creation of a postdoctoral fellowship program akin to the *Chandra* and *Hubble* fellowships. To provide a continuous pipeline of expertise in the areas of space science and technology, NVO provides the opportunity to attract undergraduates and high school students into scientific and technical careers, through internship and mentoring programs like the NSF's Research Experiences for Undergraduates. Interdisciplinary approaches that combine computer science, space science and astronomy, education, and journalism would be able to highlight NVO discoveries in unprecedented ways.

The Media: NASA space science missions and discoveries have sparked the imaginations of millions of television viewers and readers of newspapers and popular magazines. Images and discoveries of *Hubble, Mars Pathfinder, Chandra, Mars Global Surveyor, SOHO,* etc. have demonstrated their broad popular appeal. The NVO EPO will serve as a rich source of information, data, and scientific interpretation for the media at all levels. A media effort should be modeled after the successful programs at STScI and JPL.

A key issue that cuts across all potential partnerships is the need to provide training and professional development opportunities for the users of NVO. Training will help users become familiar with the capabilities of NVO interfaces, the type of data available, and the tools that will allow manipulation and analysis.

4.2.3 Coordination

The NVO EPO effort will enhance its effectiveness by coordinating with successful existing EPO efforts at NASA, NSF, and beyond. The NVO EPO effort will coordinate with the NASA OSS network of space science theme-oriented Forums and regional Broker/Facilitators. The NVO EPO effort will establish a pipeline for the scientific discoveries of NVO into the existing OSS EPO network, and also establish a structure of coordination at multiple levels, including partnerships and technical issues.   The NVO EPO effort also needs to be well coordinated with NSF to design the best approach that meets the needs of both agencies for NVO.

4.2.4 Evaluation

The NVO EPO effort (partnerships, activities, resources, visibility, and impact on users) needs to be informed by independent evaluation. These evaluation activities will serve as a resource for the NVO EPO, as well as inform the communities involved in NVO of best methods and practices that can serve as models. An independent EPO Working Group for NVO could provide the programmatic evaluation to ensure the goals of the program are being met, while the more in-depth assessment of impact may require the involvement and resources of an independent evaluation group.

4.2.5 Parallel and Independent EPO Competition for NVO

NVO must assure a funding approach for the EPO component that will lead to the most qualified and experienced alliances and partnerships to facilitate coordination with existing EPO efforts, and an early beginning of EPO activities.

5 Implementation of the NVO

Previous sections have described the technological changes that will enable a "new astronomy" and the characteristics of an NVO that can capitalize and build upon those changes to enable new and more cost effective science than would otherwise be possible.

The fundamental basis for the NVO management activities will be to recognize the science driven nature of the NVO and to maximize the community participation in the NVO effort. In general, a management structure could involve three levels. These would be structured to ensure that there is a usable and well-documented infrastructure, that the software projects are science driven, and that bulk of the funding is dispersed to well-focused science based proposals that are peer reviewed.

The highest priority is to build the archive infrastructure and well-documented protocols to access the data. These will be standards for data access that the community can rely on to build higher-level tools. These would evolve as the technology advances, but should always be backward compatible. This infrastructure should be funded via a base budget and developed and maintained by the major NVO sites.

A second level of implementation would provide opportunities to build software tools that utilize the above infrastructure. These efforts could be funded by mechanisms similar to the current Announcements of Opportunity. The resulting software products would be delivered to the NVO for wider use by the community and would follow standards defined by the NVO. It is important that these tool-building opportunities cover a wide range of possibilities and engage a large part of the community. A strong science enabling case for each software tool must be made, but they will be general user facilities that the entire community can use to do research.

Finally, there could be regular "AO's" to use the NVO. These would be more specific research projects with a well-defined goal that might include software development. (This would be similar to the current NASA ADP program). These would be much less structured in the sense of being grants and with the deliverable being a paper to a journal.

In the early phases of the NVO, the emphasis may be on the first two areas, but as the NVO infrastructure develops the balance of the funding between these three areas will evolve. A

major objective of any implementation plan is to begin providing some levels of functionality as quickly as possible through use of existing tools and services. However, the development of the NVO capabilities should pursue aggressively the full scope of the enabling information technology. The NVO must be capable of evolving as the information technology and astronomy evolve, and the new research directions and needs open up.

5.1 Organizational Requirements

Given the Charter of the Science Definition Team, we do not attempt to put forward a detailed Management Plan for the NVO. Instead, what is presented here is a brief description of some of the *key issues* that must be addressed in any implementation scheme, however the NVO is funded, and independent of the details of a management plan. It is the view of this SDT that these issues must be successfully addressed in any manifestation of the NVO if the NVO is to be a viable and long lasting component of the national and international astronomy communities.

5.1.1 Communication and Coordination

Any successful model of the NVO must clearly incorporate robust and effective connections with many entities, both within and without the astronomical community. The principal linkages that the NVO must establish from the outset are the following:

(a) **Communication with the Astronomical Community as a Whole:** This is an absolutely essential element if the NVO is to succeed, and it will be difficult to construct. This connection cannot be created by executive decision or by the forging of some cooperative agreement. Instead it must be carefully created through the establishment of trust and credibility on the part of the NVO. This in turn requires thoughtful and patient development of extensive lines of two-way communication. The community must be continuously informed of what the NVO is doing, and the NVO must continuously solicit, and clearly respond to, input from the community. The importance of making the community aware of the value of the NVO to all astronomers cannot be overemphasized.

(b) **Coordination with Existing Astronomical Data Centers:** This includes the existing NASA centers as well as the publicly funded national ground based observatories. In addition, links must be established with the major privately funded observatories, both optical and radio.

(c) **Interfacing with NASA, NSF and Other Agencies:** Different funding agencies have different objectives, cultures, and funding criteria. The NVO must be structured so that it can seamlessly accommodate the differing requirements from various funding agencies. This in turn implies a flexible and streamlined management structure that is characterized by adaptability and rapid response.

(d) **Coordination with International Efforts:** Virtual Observatory activities are underway in many other countries, and the level of effort is especially high in Europe and the UK. Since the Virtual Observatory will eventually become a worldwide entity, it is essential that the US NVO begin coordination of its activities with those in other countries at the earliest possible time. This will enable common structures and protocols to be established from the outset and will avoid expensive and unnecessary duplication of effort later in the program. The early stages of such coordination is already taking place between the US and European VO groups.

(e) Collaboration with Other Disciplines: An essential part of the NVO is its dependence on, and integration with, ITR research and development efforts. This connection is deeply embedded in all NVO design documents, and ITR activities directly related to the NVO are already underway. In addition to ITR, there are significant data archiving and management projects either planned or under development in other fields. Prominent among these are high-energy particle physics, molecular and cellular biology, and earth sciences. It is important that the NVO establish communication with researchers and institutions in these fields so that mutual advantage can be taken of developments in the different areas. Additional fields are emerging or will emerge with similar interests and problems involving data management (such as journal publication and archiving), and these groups should be incorporated into the NVO linkages as appropriate. Connections with the high-energy experimental physics community are already being established, particularly in the area of grid technology.

5.1.2   EPO Mechanisms

The enormous potential of the NVO upon education and public outreach has been described in detail in Chapter 4. In addition, that section contains a discussion of possible implementation mechanisms and recommendations for ensuring a funding approach for EPO that will lead to the most qualified and experienced alliances and partnerships.

5.1.3   Flexibility, Evolution, and a New Management Model

The NVO is a distributed federation of many different entities, with different histories and cultures, all working toward a common set of goals. In addition, the funding for the NVO will in all likelihood arise from multiple sources, and the NVO is a new concept that will evolve and mature with time. These conditions imply that a successful NVO must be characterized by a structure that accommodates evolution and adaptation, almost above all else. The NVO structure must provide coordination and communication, as well as a "single point" contact in order to respond to funding agencies and external review processes. Yet it must also incorporate the differing sizes, styles, and levels of effort of its component entities. Its functioning must be transparent and easily understood by one and all in order to obtain and retain support and credibility from its user communities.

These very diverse requirements imply that a new management model may be required for the NVO and that adoption of past management structures may not be appropriate. The successful design of this new structure is perhaps the major challenge on the road to the NVO.

5.2 Paths to Implementations

As stated in the introduction to this Chapter, the intent here is not to provide a specific implementation plan for the NVO. Many general characteristics that are felt to be essential to any successful implementation have been described in general terms in Section 5.1, especially in 5.1.3. Thus this section contains only brief summaries of some possible characterizations of implementation that have been under discussion among the members of the NVO community.

(a) Timeliness: In order for the NVO to become a viable entity, it is essential that useful and acceptable NVO functionalities be made available to the astronomy community very soon after the establishment of the NVO. Otherwise the project will lose credibility, momentum, and

relevance. The costs of delay include not only loss of interest within the community, but also the lost opportunity to capture the engagement of significant numbers of students, the loss of being an equal partner in international collaborations which are already gaining momentum, and the inevitable increased costs that result from having to reconfigure and connect databases and archives that have developed diverse structures and protocols.

Perhaps the most effective and efficient way to avoid these costs of delay is to capitalize on existing efforts and talents already extant at existing centers. Thus in addition to creating new initiatives, the NVO should serve to integrate and coordinate current activities at the existing centers. In addition, the activities described above that work toward involvement by the entire astronomical community should also be initiated as soon as possible. These could take the form of conferences, workshops, and special sessions at general meetings. NVO credibility also implies the establishment of advisory and oversight entities at a very early stage, in the form of working groups, advisory committees, and visiting committees.

(b) Interagency Coordination: The general characteristics of this topic have been described in Section 5.1.1c. The specific implementation aspect noted here is that the recent COMRAA report provides an ideal foundation for using the NVO as an outstanding example of joint support between NASA and the NSF. The required flexibility in the NVO management structure described in the previous section should incorporate procedures that will easily accommodate any differences in funding cycles, reporting requirements, and management conditions that exist between the two agencies.

(c) Long Term Stability and Structure: There are already in place several research entities funded by both NSF and NASA that have long and productive histories. A possible stable and long-term structure for the NVO could be based on the proven management models for some of these centers. This could involve management through a consortium of universities, which could in principle easily accommodate funding from different agencies. Advisory committee structure, reporting lines, and academic and financial infrastructure are all readily supplied by such a model. Alternatively, a much less structured consortium, similar to the WWW consortium, could be envisaged. However, it is important that any detailed management structure be able to incorporate the general characteristics outlined in Section 5.1, and be robust enough to protect the long-term health and functionality of the NVO from short-term funding.

6 Summary and Recommendations

6.1 General Considerations and Principles

We believe that the case for the NVO is compelling-and it is growing more so in time as the quantity, quality, and complexity of astronomical data increase. Echoing the recommendation of the Decadal Report, we see the creation of the NVO as a high priority for American astronomy and as an initiative guaranteed to produce manifold scientific returns over the years to come. The NVO is naturally a part of a broader international enterprise, leading to a truly global Virtual Observatory.

The NVO initiative represents a natural field of cooperation between the NSF and NASA, in the spirit of the COMRAA report. And while these two agencies are clearly in leadership roles

for the NVO, involvement of other federal agencies (e.g., DOE or DARPA) may be both viable and desirable.

Within the astronomical community itself there is an unprecedented range of participation. The NVO will transcend the boundaries of wavelength regimes and the traditional agency domains (e.g., ground-based vs. space-based). The constituencies of the NVO also include computer science and information technology professionals and scientists from other fields as well (e.g., statistics). Interdisciplinary efforts and mutually beneficial partnerships are crucial for the success of the NVO, and they should be actively pursued and encouraged, through, for example, interdisciplinary grant programs.

The existing astronomical data centers, together with the numerous groups involved in the creation and exploration of massive data sets and relevant information technologies represent a strong foundation for building of the NVO. The existing work conducted within the NASA centers can be directed and augmented to serve the NVO goals. However, there is an urgent need for the development and implementation of archives for major ground-based observatories. They will be equally important pillars of the NVO, alongside the already existing NASA centers and other major archives.

In addition to the technology and infrastructure development costs, a substantial fraction of the available funding should be dedicated specifically for the novel scientific investigations made possible by the NVO framework. Early NVO enabled science would serve both as a community motivator and as a testbed. A rapid development of NVO prototype services and functionalities will both stimulate and attract a broader astronomy community and will provide a critical early feedback in the development and evolution of the NVO. These services will represent the first realizations of the scientific potential inherent in the NVO concept, and engage the broad astronomy community in its development. Both the NSF and NASA can ensure that science remains the driver and the beneficiary of the NVO by funding competitive, cutting edge, NVO-based projects.

A substantial portion of the initial NVO science funding should be oriented towards the younger scientists: students and postdocs who will grow to become the new generation of leaders and experts making the most of this new field. This may be accomplished through PI grant programs and through dedicated fellowship programs that will ensure a long term growth and success of the NVO and an information-rich astronomy of the future.

The NVO has the potential of becoming an unprecedented vehicle of science and technology education in the broadest sense. The EPO component should be built in from the start, with support for strong partnerships involving scientists, NVO developers, and education and science popularization professionals. Because of the rich opportunity offered by NVO for Education and Public Outreach, there should be a parallel and independent competition for NVO EPO funding.

While there is already a healthy momentum behind the creation of the NVO, a stable stream of adequate funding is necessary if this concept is to become a reality and fulfill its scientific promise. The NVO should be seen as a source of great new opportunities in science, technology, and education, with resulting long-term benefits to the nation.

6.2 Specific Recommendations

Given these general principles and desiderata, we make the following near-term recommendations:

1. The NSF and NASA should form a task force to define the appropriate multi-agency management and funding structure for the NVO. Once such a structure is in place, a detailed implementation plan should be developed.

2. In parallel with this task force, and in order to assure continuity, maintain momentum, and lead towards the full NVO, we recommend that the NSF and NASA jointly appoint a successor to this Team, with a mandate to:

   - Advise the agencies on all issues pertaining to the NVO;

   - Coordinate the activities of various participating groups;

   - Develop, refine and modify the strategy and the roadmap towards the NVO as it evolves;

   - Continue to educate and involve the broader astronomical community in this process.

We further recommend that there be specific working groups, which may operate under the umbrella of the SDT successor on the subjects of international cooperation and coordination, technical issues (standards, architecture, etc.), EPO, and possibly others.

3. The NSF and NASA should designate a portion of the funding from their existing research and analysis programs specifically for the development of the NVO and NVO-based science. Funding of the early NVO demonstration prototypes should be a priority. A more extensive, dedicated NVO funding stream should be built in the future budgeting process.

4. We recommend the establishment of NVO postdoctoral, and graduate and undergraduate fellowships.

5. NVO must include a funding approach for the EPO component that will lead to the most qualified and experienced alliances and partnerships to facilitate coordination with existing EPO efforts, and that will ensure an early beginning of EPO activities.

6.3 An Outline of the Development Process and Timeline

We see a phased development of the NVO as follows:

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.

Appendix E gives an example of a plausible budget for the development of the NVO for the period of ten years.

6.4 Conclusions: NVO as a New Research Environment for the Astronomy of 21st Century

The progress in science, astronomy included, tends to be gradual, punctuated with bursts of creative growth, which follow introduction of major new technologies. In astronomy, such an episode happened in the early 1960s, with the advent of radio- and x-ray astronomy; this resulted in the discoveries of quasars, pulsars, the cosmic microwave background. Again in the mid-1990s another episode occurred, with the discoveries of forming and evolving galaxies at high redshifts, extrasolar planets, brown dwarfs, CMBR fluctuations, and many more, thanks in large part to advances in telescopes and detector technologies. In both cases, new data-gathering technologies ignited small scientific revolutions.

We are standing on the threshold of a new era of discovery in astronomy. Information and detector technologies are giving us data sets many orders of magnitude larger, richer, and more complex than anything we ever had before, and we have barely started to explore them. The NVO will be our mechanism in making possible the next fundamental advances, and become *an engine of discovery for astronomy.*

Appendix A. The Existing Efforts: Foundations of the NVO

In assessing the current state of North American astronomy, the following resources are already in place to support the emerging NVO:

Data Centers and Supercomputer Centers. Some tens of terabytes of data products (catalogs, images, and spectra) already exist for various space missions, public telescopes, and surveys; this will expand to a petabyte or more of data by the end of the decade. Archive and data analysis capabilities exist at the major NASA centers (STScI-MAST, IPAC-IRSA, GSFC-HEASARC, and CXC) and at the CADC (Canada); many smaller or more focused archives exist as well. Supercomputer centers such as the SDSC and NCSA are available for addressing large-scale computational problems. A high performance national networking infrastructure is already in place.

Astronomical Information Services. Information services such as the ADS, NED, and SIMBAD exist for name resolution and cross-referencing of galactic and extragalactic objects, and are providing increasingly sophisticated levels of interlinking between bibliographic information, the refereed and preprint literature, and the archival data centers.

Data Analysis Software. Various software packages such as AIPS, AIPS++, CIAO, IRAF, IDL, FTOOLS, SkyView, etc., exist for the general analysis of astronomical data. The development of sophisticated software for large scale data mining is still in its infancy, although new initiatives such as the NPACI-sponsored Digital Sky and the IPAC Infrared Science Archive are showing the potential of such facilities and have prototyped the technology required to correlate and mine such data archives.

Although these resources are significant, anyone who has tried to perform multi-wavelength data analysis or large scale statistical studies combining several different catalogs, with the data involved being available from widely distributed and dissimilar archives, will appreciate how far we have to go to implement the vision of the NVO. Ground-based O/IR and radio data need to be pipeline-processed and archived routinely as space-based data are now. Standards and protocols need to be developed to allow widely distributed archives to interoperate and exchange data. Astronomical data analysis software needs to evolve to be able to access data in distributed multi-wavelength archives as easily as local datasets are accessed now. New algorithms, applications, and toolkits need to be developed to mine multi-Terabyte data archives. Supercomputer-class computational systems need to be developed to enable large-scale statistical studies of massive, multi-wavelength distributed data archives. The data, software, and computational resources need to be interconnected at the highest available network bandwidths.

A.1 Existing Astronomical Data Centers and Archives

A.1.1 NASA Data Centers and Archives

NASA's Astrophysics Data and Information Services (ADIS) provide access to the data, associated software and expertise in its use from NASA's astrophysics missions, with the prime goal of maximizing the use of these data by the world community. Immediate on-line access to the entire array of space-based astrophysics data is now taken for granted, and multi-mission data analysis is routine. Web-based software tools allow users to explore and visualize a rich environment of data holdings, retrieve data, software, and associated publications and instantly find comprehensive information on a wide range of objects. A very successful education and public outreach program engages the general public and educators. Reuse of software across missions is now common and has resulted in cost savings to the NASA program (e.g., STScI's OPUS pipeline processing software and Spike planning and scheduling tools have been adopted by several mission science centers).

NASA's Office of Space Science has embraced the systematic archiving of data from space astrophysics missions for nearly two decades. This enlightened vision for comprehensive data management has culminated in an astrophysics data system composed of primary science archive research centers (SARCs): a high-energy SARC, an optical/UV SARC, and an infrared SARC. The High Energy Astrophysics SARC (HEASARC) is located at NASA/Goddard Space Flight Center, the optical/UV SARC, known as the Multi-mission Archive at Space Telescope (MAST), is located at the Space Telescope Science Institute in Baltimore, and the infrared SARC, known as the Infrared Science Archive (IRSA), is located at the Infrared Processing and Analysis Center (IPAC) at Caltech.

The National Space Science Data Center (NSSDC) provides permanent archive services to the three SARCs and provides direct archival services for the COBE and IRAS mission data sets. Active missions such as SIRTF, HST, and Chandra process and manage their data directly at the associated science operations center (SIRTF Science Center, Space Telescope Science Institute, Chandra X-Ray Center). Active missions coordinate with the relevant SARC to assure access to their data during the mission lifetime and arrange for long-term access after the mission is over. In some cases active missions contract directly with one of the SARCs to provide archive and data distribution services (e.g., the FUSE mission data are processed at Johns Hopkins University and archived at MAST/STScI).

The current relationships between the NASA mission centers and research teams, the NASA data archive centers, the NASA information/integration services, and the user communities are laid out in Figure 2.
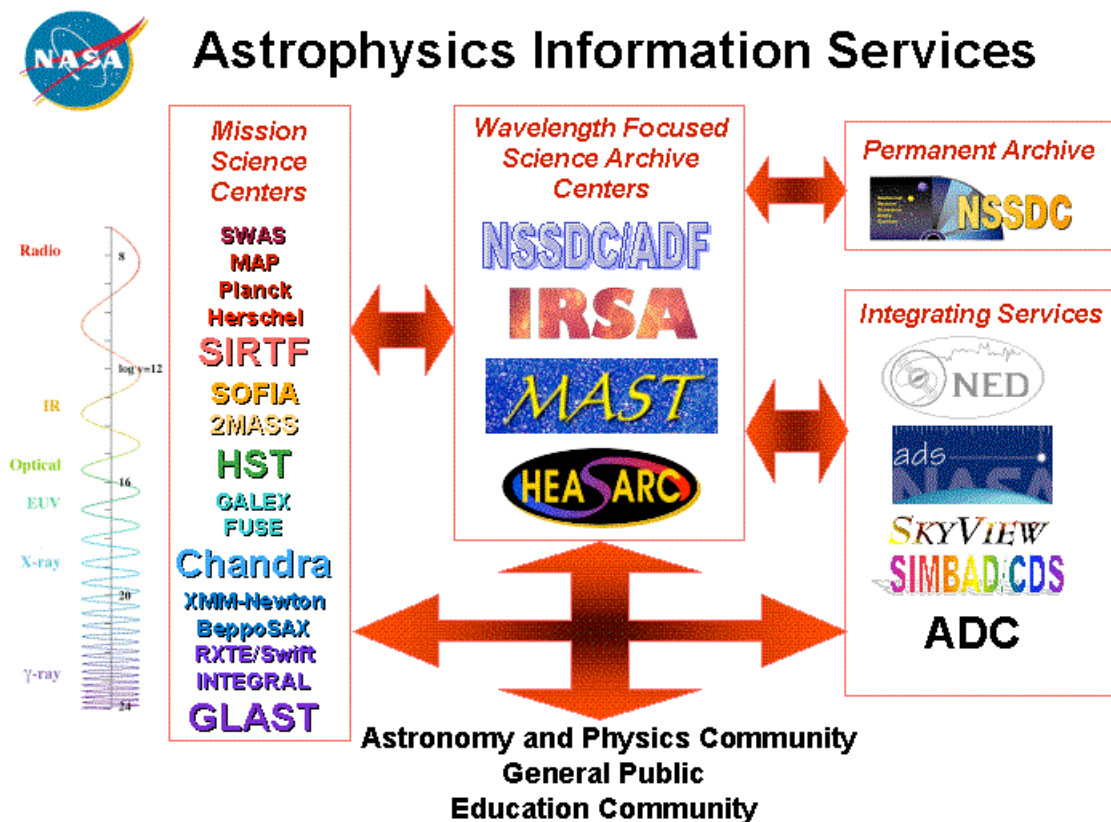
Figure 2. NASA's Astrophysics Information Services

A brief summary of the roles of the mission science centers, data archive facilities, and "integrating" information services is provided below:

Wavelength-Focused Science Archive Centers:

High Energy Astrophysics Science Archive Research Center: The HEASARC, located at the GSFC, enables archival research using Gamma-ray and X-ray astronomical data (http://heasarc.gsfc.nasa.gov/). The HEASARC was established by NASA in November 1990 as the first of three wavelength specific science archive research centers. Online access to astrophysics data was pioneered by the HEASARC and its entire data and software holdings are available for immediate download via the Internet. The Skyview facility that provides images of the sky in any waveband is hosted by the HEASARC. The HEASARC has helped define standards for data, software and archive interoperability that have resulted in substantial cost savings and improved data accessibility. The contents of the HEASARC's data holding totaled 2.2 Terabytes as of the end of 2000, a 33% increase over the size at

the end of 1999. The 3 currently operating missions that are supplying data to the HEASARC are RXTE, BeppoSAX, and XMM-Newton. Guest observer facilities for these missions are co-located with the HEASARC. The HEASARC is now a partnership between GSFC and SAO, with Chandra data accessible via the HEASARC interfaces. The bulk of the remainder of the HEASARC Archive consists of data from four missions that were operational in the 1990s: ASCA, CGRO, ROSAT, and EUVE, and from four missions that were operational in the 1980s: EXOSAT, Ginga, Einstein, and HEAO-1. There are also small volumes of data from about 10 other missions. Finally, in addition to data and calibration files associated with specific missions, the HEASARC archive also contains has 8.2 GB of multi-mission software such as the HEAsoft package. For the future the HEASARC plans to archive data from the following approved missions: INTEGRAL, Swift, Astro-E2 and GLAST.

Infra-Red Science Archive: The NASA/IPAC InfraRed Science Archive (IRSA), located in Pasadena, is a project focused on providing software and Internet services to facilitate astronomical discoveries, support the production of new astronomical data products, and to plan future observations utilizing the data archives from infrared astrophysics missions supported at IPAC: the Two Micron All-Sky Survey (2MASS), the Space InfraRed Telescope Facility (SIRTF), the Infrared Space Observatory (ISO), the Midcourse Space Experiment (MSX), and the Infrared Astronomical Satellite (IRAS). IRSA's data holdings total nearly 15 TB of catalog, images, and spectra. Since January 2001, IRSA has processed over 1,800,000 data requests from over 30,000 independent hosts, and delivered nearly 450 GB of data. While many capabilities are driven by the requirements of the very large archive currently being generated by the 2MASS survey, the database and software architecture are designed to support the archive access needs of each of these projects. The software tools and services are referred to as the IRSA Information System, or ISIS. ISIS uses a distributed processing model, where data and services can be accessed independent of location and where all user interface (http://irsa.ipac.caltech.edu/) and communications functions (on the user side) are handled by standard Web browsers. Like all the NASA data archives, IRSA is successfully leveraging resources at other NASA centers and elsewhere to enable new functionality, while at the same time supporting and collaborating with these remote archive systems.

Multi-mission Archive at STScI: The Multi-mission Archive at Space Telescope Science Institute (MAST) is NASA's optical/UV science archive center. MAST, established in October 1997, builds upon the infrastructure developed for the HST archive but expands the service to support 12 additional missions, including FUSE (active), IUE, EUVE, Copernicus, three ASTRO missions, three ORFEUS missions, the Digitized Sky Survey and the VLA FIRST survey. Future additions will include the Sloan Digital Sky Survey, the second generation Guide Star Catalog, GALEX, and CHIPS. Partial funding for the ground-based surveys (VLA FIRST, DSS, SDSS, and GSC 2.0) archived at MAST comes from sources external to NASA. There are presently 3,300 unique MAST users, including professional astronomers, graduate students, and educators. The primary user interface to MAST is via the website http://archive.stsci.edu/mast.html. Users can query MAST's on-line database to explore its data holdings, cross-correlate catalogs of positions with the holdings, request data, read on-line mission documentation, download mission data analysis software, access FAQ pages, and link to other NASA data centers.

National Space Science Data Center / Astrophysics Data Facility: The NSSDC's ADF, based at the Goddard Space Flight Center (GSFC), provides active archiving of selected long wavelength data, specifically data from IRAS, COBE and SWAS. For IRAS and SWAS, NSSDC/ADF provides web-based interfaces which link to data files of individual data sets (see http://hypatia.gsfc.nasa.gov/adf/adf.html). For COBE, the ADF provides this plus a broader range of tools for browsing of the data. The ADF will provide support for data from the Microwave Anisotropy Probe (MAP) mission similar to that provided for COBE.

Mission Science Centers:

The mission science centers are generally responsible for all phases of a mission's science operations from overseeing the peer-reviewed proposal selection process, to the execution of the observations, to the calibration of the data, and ultimately, to the dissemination of the data to the professional community and the general public.

Space Telescope Science Institute (STScI): STScI is the science center for the Hubble Space Telescope mission (http://www.stsci.edu). The HST was the first of NASA's Great Observatories and was launched in April 1990. The Institute was established in 1981 and is located in Baltimore, MD. STScI is also the home site of NASA's optical/UV science archive facility (MAST). The HST archive presently contains over 7 TB of data and is growing with ~100 new science exposures every day. The next HST servicing mission will see the installation of the Advanced Camera for Surveys and the re-activation of the Near Infrared Camera and Multi-Object Spectrograph. STScI has been selected as the science center for the Next Generation Space Telescope.

Chandra X-ray Center (CXC): The Chandra X-Ray Observatory is the latest of NASA's Great Observatories: a high-resolution imaging and spectrographic telescope operating in the X-ray part of the electro-magnetic spectrum. Chandra was launched on July 24, 1999. The Chandra Data Archive (CDA) is part of the Chandra X-Ray Observatory Science Center (CXC; http://asc.harvard.edu/) which is operated for NASA by the Smithsonian Astrophysical Observatory in Cambridge, MA. The current holdings of the CDA amount to several million data products with a total volume of 2 TB, in addition to an extensive collection of databases that hold mission information and metadata on the data products. The Chandra archive volume is expected to expand by almost 1 TB per year of active mission.

SIRTF Science Center: The Space InfraRed Telescope Facility (SIRTF) is the fourth and final element in NASA's family of Great Observatories. SIRTF consists of a 0.85-meter telescope and three cryogenically cooled science instruments capable of performing imaging and spectroscopy in the 3 - 180 micron wavelength range. Incorporating the latest in large-format infrared detector arrays, SIRTF offers orders-of-magnitude improvements in capability over existing programs. While SIRTF's mission lifetime requirement remains 2.5 years, recent programmatic and engineering developments have brought a 5-year cryogenic mission within reach. The SIRTF Science Center (http://sirtf.caltech.edu/) is co-located with the Infrared Processing and Analysis Center (IPAC) on the campus of the California Institute of Technology.

GLAST Science Support Center (SSC): The Gamma ray Large Area Space Telescope (GLAST) is the next major NASA Gamma ray mission and will be launched in 2006. A science support center has been established at NASA-GSFC, which builds upon a similar center established in 1990 to support the Compton Gamma Ray Observatory. The GLAST SSC will provide processing, archiving, user support, mission planning and analysis software for the GLAST mission. The GLAST SSC is co-located with the HEASARC.

SOHO Archive: The Solar Heliospheric Observatory (SOHO) archive at GSFC contains the data from the suite of 12 instruments on the spacecraft. These instruments comprise a total solar observing package that has provided nearly continuous data since early 1996. (http://umbra.nascom.nasa.gov/sdac.html)

TRACE Data Center: The Transition Region and Coronal Explorer mission is continually producing an open on-line archive of the data, with no restrictions on access. This innovative policy, coupled with the ease of data access, has substantially accelerated the research process and significantly increased the scientific productivity of the mission. (http://vestige.lmsal.com/TRACE/)

NASA Data Center Services:

Several of the NASA data centers have already implemented NVO-like services. For example, HEASARC and MAST share responsibility for the EUVE mission data, which physically resides at HEASARC but is available to users through both HEASARC and MAST interfaces. Users need not be aware of the physical location of the data. IRSA provides access to the full 2MASS data archive, but the data are physically located on mass-storage media at the San Diego Supercomputer Center, again transparent to the end-user. CXC has pioneered the development of generic data models that allow computer programs to access a wide variety of basic data types and formats without the need for custom software revisions. This approach, along with standardization of the associated descriptive information about data-the metadata-is an essential element of the nascent NVO.

Similarly, HEASARC's Astrobrowse, SkyView, and Browse tools, MAST's Spectral Scrapbook and Starview-II, and IRSA's OASIS provide basic cross-archive data location and integration services. All are important pathfinders for future NVO capabilities and services. For solar, the Solar Data Analysis Center (SDAC) at GSFC stores the data from past NASA solar missions (e.g. Yohkoh, SMM) as well as several major ground-based observatories. SDAC also provides a number of data analysis tools to the solar community.

Integrating Services:

Astrophysics Data System Abstract Service: The NASA Astrophysics Data System Abstract Service (http://adswww.harvard.edu) has become a key component of astronomical research. Located at the Center for Astrophysics in Cambridge, ADS provides bibliographic information daily to a majority of astronomical researchers worldwide. The use of the ADS is much larger than the use of all traditional astronomical libraries in the world combined. Astronomy is unique in that it already has a fully functional data resource, where several of the most important data sources exist on-line and inter-operate nearly seamlessly. The ADS and the Strasbourg Data Center (CDS) form the core of this resource.

Astronomical Data Center: The Astronomical Data Center, located within the NSSDC at GSFC, provides increasing access, display and utilization functionalities to the growing number of astronomical catalogs and journal tables that it manages. The ADC (http://adc.gsfc.nasa.gov/) has developed significant access and display software tools (e.g., IMPReSS, CatsEye, Viewer) in recent years. In the past two-three years, ADC has played a lead role in the application of XML (eXtensible Markup Language) technology to NASA's needs in astrophysics data management and has, in particular, developed XML-based tools for the automated ingestion of catalogs and tables and for facilitating the retrievability of their contents. ADC has an extensive and complimentary collaboration with Europe's CDS. For example, ADC presently is a mirror site for the CDS *VizieR* software.

NASA Extragalactic Database: NED (http://ned.ipac.caltech.edu) is an Internet-based research facility that supports on-going and up-coming NASA astrophysics missions, as well as observatories, scientists, and educators in the planning, execution and publishing of multi-wavelength research into extragalactic astronomy and cosmology. NED provides up-to-date and on-line electronic access to a comprehensive and detailed database of nearly 5 million galaxies, quasars and extragalactic radio, x-ray and infrared sources. The main database contains accurate positions, redshifts, robust cross-identifications, and associated physical quantities (including multi-wavelength photometry, images, and other basic data, derived from space-based and ground-based surveys as well as from targeted observations.) The information is rapidly and easily retrieved by the user, and all tightly linked to the peer-reviewed, published literature. These quantities can be queried in both simple and complex ways to support and facilitate nearly all aspects of observational and archival extragalactic research. NED gives its users the ability to access globally distributed datasets that are related to specific objects or positions on the sky as indexed by NED.

NASA Data Center Management Structure:

For the past several years NASA's astrophysics data centers have been working together to improve and expand their services to the community, focusing particularly on providing common front-end services and transparent access to each other's data holdings. The goal is to minimize duplication of effort, build upon each other's strengths, and enable the science community to more easily utilize our complementary services. To this end, an Astrophysics Data Executive Committee (ADEC) made up of representatives from each of the NASA data centers and related information services, provides a forum to coordinate and steer the activities. The ADEC consists of a senior representative from each data or service center (ADS, HEASARC, IRSA, MAST, NED, NSSDC/ADC), and a senior representative from each of the largest active mission operations centers (CXC, GLAST, SIRTF, STScI/HST). The ADEC is led by a chairperson elected from among the representatives, with a one-year term of office. The chair acts as the coordinator and consensus builder and act as an advocate for the ADIS within the community. An operating plan is produced by the ADEC annually and submitted to NASA HQ. This plan highlights achievements from the past year, plans for the future years and coordination of efforts to provide a uniform approach to NASA's ADIS. NASA OSS is now constituting a Science Archives Working Group (SAWG), reporting to the Origins and Structure and Evolution of the Universe Subcommittees, to provide feedback and guidance on matters that require community review and consensus.

A.1.2 Ground-Based Observatories

Ground-based observatories in the United States, both public and private, have until recently provided only the most basic archival services, if any at all. The National Optical Astronomy Observatories have for some years maintained a "Save-the-Bits" archive, but this is primarily a save-store that guards against data loss. It lacks an on-line presence, and does not provide sufficient documentation to allow users other than the original observer to be able to make use of the data. The National Radio Astronomy Observatory also has a data archive, but like NOAO it is primarily a data safe-store. An on-line catalog allows astronomers to locate data sets of potential interest, but these must be requested through a largely manual process (requiring the user to contact the original observers to obtain permission to use the data). The result of a data request is a series of uncalibrated visibility data sets. The National Center for Supercomputing Applications created the Astronomy Digital Image Library (ADIL) several years ago as a way to provide on-line access to processed images. ADIL has been somewhat successful in radio astronomy, especially in support of mm-wave observations from BIMA.

Fortunately, the public ground-based observatories are now beginning to develop full-function electronic archives. NOAO's efforts are being catalyzed through, e.g., the Deep-Wide Survey project and other dedicated surveys being performed explicitly to provide wide public access. NRAO is revamping its approach to data

management, at least partly motivated by the need to plan for the tremendous data volumes that will ensue from the ALMA telescope.

With the exception of the Sloan Digital Sky Survey (funded in the majority by private resources) private observatories have done little in the way of archiving and data management. The SDSS, however, is making all of its data publicly available under an agreement reached with the National Science Foundation. Our expectation is that as the NVO matures, the sponsors of private observatories will begin to understand the benefits of making their data more widely available (at the same time as data storage costs continue to decrease). In preliminary discussions with the leadership of the QUEST survey project, for example, Yale University indicated substantial interest in participating in the NVO.

In solar, there are several ground-based observatories that are regularly archiving their data. The National Solar Observatory (NSO) has an on-line Digital Library containing 25 years of data from the Kitt Peak facilities. This library is about to expand with the addition of SOLIS and GONG+ data. Other observatories with considerable archives include the Big Bear Solar Observatory, the High-Altitude Observatory, the Mees Solar Observatory, Mt. Wilson, the Wilcox Solar Observatory, and California State University at Northridge.

A.1.3 International Data Centers

CDS: Perhaps the foremost international data center is the CDS-Centre de Données astronomiques de Strasbourg-hosted by the Strasbourg Astronomical Observatory in France (http://cdsweb.u-strasbg.fr/). The CDS is dedicated to the collection and worldwide distribution of astronomical data and related information. The CDS hosts the **SIMBAD astronomical database,** the world reference database for the identification of astronomical objects.

The goals of the CDS are to:

- Collect all of the useful information concerning astronomical objects that is available in computerized form: observational data produced by observatories around the world, on the ground or in space;

- Upgrade these data by critical evaluations and comparisons;

- Distribute the results to the astronomical community; and

- Conduct research, using these data.

The CDS plays, or has played, a part in most of the major astronomical space missions: generating guide star catalogues (EXOSAT, IRAS, Hipparcos, HST, ISO, SAX), helping to identify observed sources (Hipparcos, Tycho, ROSAT, SIGMA), or to organize access to the archives (IUE), etc. CDS contributes to the XMM Survey Science Center, with the High-Energy group of Strasbourg astronomical observatory. The CDS is cooperating with ESA (e.g. transfer of ESIS Catalogue Browser to CDS: the VizieR project), and with NASA: in particular, CDS hosts a mirror copy of the Astrophysics Data System (ADS) project, while ADS hosts a mirror copy of SIMBAD at SAO. CDS also contributes to the AstroBrowse project. CDS also hosts the European mirrors of the journals of the American Astronomical Society (AAS).

CADC: The Canadian Astronomy Data Centre is located at the Dominion Astrophysical Observatory in Victoria, B.C., Canada, part of the Herzberg Institute of Astrophysics under the National Research Council of Canada. The CADC was established in 1986 as one of three worldwide distribution centers for data from the Hubble Space Telescope. Funding is provided by the Canadian Space Agency. The CADC is also responsible for archiving data from the Canada France Hawaii Telescope.

ESO and ST-ECF: The ESO/ST-ECF Science Archive is a joint collaboration of the European Southern Observatory and the Space Telescope-European Coordinating Facility. The ST-ECF (http://ecf.hq.eso.org/ecf/) was established in 1984 jointly by the European Space Agency and the European Southern Observatory and is located at the ESO headquarters at Garching near Munich. The ST-ECF staff supports the European astronomical community in exploiting the research opportunities provided by the earth-orbiting Hubble Space Telescope. The ST-ECF provides detailed technical information about the HST and its *science instruments*, supports European astronomers in the preparation of HST observing proposals coordinates the development of computer software tuned to the specific data analysis needs of HST users, operates and maintains an archive of all the scientific data collected by HST, and acts as a European center for associated meetings and workshops. The ESO archive provides access to scientists from ESO member states and Chile to data from the various ESO telescopes and instruments.

A.1.4 Major Survey Archives∗

Radio: The radio sky has been surveyed a number of times over the years at different frequencies and increasing sensitivities. One of the most significant in recent years was performed using the Very Large Array. The NVSS (NRAO VLA Sky Survey) is covering the sky north of $\delta > -40°$ at 1.4 GHz and is now 99.9% complete. Over 1.8×10^6 sources with a limiting sensitivity of 2.5 mJy have been identified. The catalog and over 2TB of image data are all available on-line from NRAO and NCSA. See http://www.cv.nrao.edu/nvss/.

A second VLA survey is also well underway (80% complete) of the 10,000 deg^2 around the North Galactic Cap in order to complement the optical Sloan Digital Sky Survey. This FIRST survey (Faint Images of the Radio Sky at 20-cm) is also being performed at 1.4 GHz but with greater resolution (1.8") and to a deeper limiting sensitivity (1 mJy). These data are also being processed and the image data and derived catalogs being made publicly available on web sites at STScI, LLNL, and NRAO. See http://sundog.stsci.edu/top.html.

The southern hemisphere counterpart to the NVSS is being carried out by the Molonglo Observatory Synthesis Telescope, which is observing the sky south of $\delta < -30°$ at 843 MHz. The SUMSS survey (Sydney University Molonglo Sky Survey) has observed approximately 53% of the area and data covering 1500 deg^2 has been released on-line at the University of Sydney using the same data format adopted by the NVSS. See http://www.astrop.physics.usyd.edu.au/SUMSS/.

Infrared: The 2MASS survey (2-Micron All Sky Survey) was the first project to perform a ground-based all-sky survey using digital detectors. Two dedicated 1.3m telescopes at Mt. Hopkins and CTIO have been imaging (1" resolution) the entire sky in the infrared J (1.25μm), H (1.65μm) and Ks (2.17μm) bands. Complete coverage was recently achieved and 50% of the data have already been released by IPAC. There will ultimately be over 10TB of image data and a catalog with about 10^8 sources. See http:// www.ipac.caltech.edu/2mass/.

In the southern hemisphere the ESO 1m telescope has performed a slightly deeper survey than 2MASS in the Ig (0.8μm), J (1.25μm), and Ks (2.16μm) bands. The DENIS survey (Deep Near Infrared Survey of the Southern Sky) completed observations of the southern hemisphere in September 2001. Although only 2% of the data has been publicly released at this time, eventually 3TB of images and a source catalog of 5×10^8 will be available on-line from the CDS. See http://cdsweb.u-strasbg.fr/denis.html.

Optical: Schmidt telescopes have been performing surveys for several decades. Until recently, these were the only practical means of obtaining all-sky images to a reasonable depth with a reasonable resolution. The Palomar observatory in the northern hemisphere, and the UK and ESO Schmidt telescopes in the southern hemisphere, have conducted a number of surveys.

These surveys were distributed to the community on either glass plate or film, and a number of different scanning machines have been used to digitize most of these to date. Each of these scanning machines have different characteristics, project goals and processing algorithms such that one can use their different results as quality control checks to validate each other. Most of these groups have placed the available images and derived object catalogs on-line.

The Sloan Digital Sky Survey (SDSS, http://archive.stsci.edu/sdss/index.html and http://www.sdss.org/) is designed to image approximately π steradians centered on the northern galactic cap in 5 bandpasses (u', g', r', i', and z') and perform follow-up spectroscopy to determine one million redshifts. This is being carried out with a dedicated 2.5m telescope at Apache Point Observatory. Routine operations started in 2000 and will continue for at least 5 years. The scientific potential of this survey has already been demonstrated with an analysis of the test data taken during the commissioning. This has included the discovery of several high-z quasars and methane dwarfs. Even though only 10^4 deg^2 will be covered by the SDSS, this is the type of survey that will be the successor to the photographic surveys with deeper imaging and higher resolution pixels (0.4" pixels). When completed, there will be approximately 15TB of calibrated images available and a catalog with 10^8 objects. The data will be incrementally released to the community through MAST (the Multi-mission Archive at STScI), following a proprietary period for consortium members. The spectroscopic component of the SDSS is a natural extension of previous and current

---

∗ Adapted from B. McLean, in ASP Conf. Ser. 225, 103 (2001)

galaxy redshift surveys that have been important in our understanding of large-scale structure (CfA, Las Campanas, 2dF, 6dF, etc.).

On-line archives and catalogs are now integral activities for all major surveys. There are an estimated 50TB of ground-based (radio-IR-optical) survey data available and an atmosphere of cooperation between the different groups so that the virtual observatory concept has a solid base to build upon. There is also a large amount of non-survey data available in observatory archives and data centers around the world. Improving connectivity and accessibility of all these data will enrich the entire community who will use this International Virtual Observatory.

Table 3: An incomplete list of major data collections as of early 2001.

Name	Data Type	Spectral Coverage	Size	No. of Observations	No. of Sources	Responsible Organization
IUE	S	1200-3350 Å	600 GB	104,000	10,000	STScI/MAST
EUVE	S,I	70-760 Å	61 GB	1150	400	STScI/MAST
Copernicus	S	900-3150 Å	1 GB	687,000 scans	551	STScI/MAST
ASTRO UIT	I	1200-3300 Å	56 GB	1600	200	STScI/MAST
ASTRO HUT	S	825-1850 Å	0.6 GB	500	300	STScI/MAST
ASTRO WUPPE	P	1400-3200 Å	0.1 GB	400	200	STScI/MAST
ORFEUS BEFS	S	900-1200 Å	4 GB	317	108	STScI/MAST
ORFEUS TUES	S	900-1400 Å	0.6 GB	239	62	STScI/MAST
ORFEUS IMAPS	S	950-1150 Å	0.3 GB	600	10	STScI/MAST
HST	S,I,P	1100-25000 Å	>7 TB	>230,000	>20,000	STScI/MAST
FUSE	S	905-1190 Å	>25 GB	>1160	>800	STScI/MAST
DSS	I	4400-7000 Å	465 GB	4780 plates		STScI/MAST
VLA FIRST	I	20 cm	110 GB	15,000	720,000	STScI/MAST
GSC II	I,C	4000-8500 Å	2 TB	3500 plates	2×10^9	STScI/MAST
SDSS Early Release	I,S,C	3600-9200 Å	1 TB	15,000	10^7	STScI/FNAL/JHU
SDSS	I,S,C	3600-9200 Å	20 TB	1×10^6	3×10^8	FNAL/JHU/STScI
Ariel V	P,C	0.3-40 keV	0.1 GB	250		GSFC/HEASARC
ASCA	S,I,P,C	0.4-10 keV	535 GB	3833		GSFC/HEASARC
BBXRT	S,P	0.3-12 keV	2.1 GB	157		GSFC/HEASARC
Beppo-SAX	S,I,P	0.1-300 keV	50.3 GB	3462		GSFC/HEASARC
CGRO	C	30 keV - 30 GeV	279 GB	~10,000		GSFC/HEASARC
COS-B	I	2 keV - 5 GeV	0.1 GB	252		GSFC/HEASARC
Copernicus UCLXE	R	0.5-10 keV	0.4 GB	6555		GSFC/HEASARC

Name	Data Type	Spectral Coverage	Size	No. of Observations	No. of Sources	Responsible Organization
Einstein (HEAO-2)	S,I,P,C	0.2-4.5 keV	15 GB	~6000		GSFC/HEASARC
EUVE	S,I	70-760 Å	61 GB	1150	400	GSFC/HEASARC
EXOSAT	S,I,P,C	0.05-20 keV	107 GB	6614		GSFC/HEASARC
Ginga (Astro-C)	S,P	1-500 keV	19.7 GB	11,673		GSFC/HEASARC
HEAO-1	S,I,P,C	0.2 keV - 10 MeV	9.5 GB	~10,000		GSFC/HEASARC
OSO-8	P	0.15 keV - 1 MeV	6.4 GB	~2500		GSFC/HEASARC
ROSAT	S,I,P,C	0.1-2.5 keV	181 GB	15,000		GSFC/HEASARC
RXTE	S,P,C	2-250 keV	883 GB	24,561		GSFC/HEASARC
SAS-2	I	20 MeV - 1 GeV	0.1GB	81		GSFC/HEASARC
SAS-3	R	0.1-60 keV	7.2 GB	321		GSFC/HEASARC
Vela 5B	P	3-750 keV	0.1 GB	268		GSFC/HEASARC
Uhuru (SAS-1)	C	2-20 keV	0.1 GB	339		GSFC/HEASARC
XMM	S,I,P,C	.1-15 keV	1.5 GB	3		GSFC/HEASARC
HEASARC Catalogs	C	mostly > 1 keV	2 GB	~1 x 10^6		GSFC/HEASARC
NEAT/SkyMorph	I,C	7,000 A	2.5 TB	80,000	4×10^8	JPL-GSFC/HEASARC
WENSS	I	325 MHz	0.5 GB	500		HEASARC/SkyV.
SUMSS	I	843 MHz	1.0 GB	108		HEASARC/SkyV.
GB6	I	4850 MHz	2.5 GB	616		HEASARC/SkyV.
IRAS	I,C	12,25,60,100 _	10 GB	1720	1×10^6	IPAC/IRSA
MSX	I,C	8.3 _	200 GB	1590	330,000	IPAC/IRSA
2MASS	I,C	1.25, 1.65, 2.17 _	15 TB	4×10^6	3×10^8	IPAC/IRSA
IRAS	I,C	12,25,60,100 _		1720	1×10^6	GSFC/NSSDC
COBE	S,I,P	1.25 _ - 10 mm	28 GB	3×10^9	n/a	GSFC/NSSDC
SWAS	S	487-557 GHz	1.1 GB	6000	77	GSFC/NSSDC
Chandra	S,I,P,C	0.3-8 keV	1 TB	3000	500	SAO/CXC
Mosaic North	I		10 TB	72,000		NOAO
NDWFS	I,C		300 GB			NOAO
NOAO Surveys	I,C		2 TB			NOAO

Name	Data Type	Spectral Coverage	Size	No. of Observations	No. of Sources	Responsible Organization
NVSS	I,C					NRAO
FIRST	I,C	20 cm	110 GB	15,000	720,000	NRAO
VLA	I,R	0.7-400 cm	2.5 TB	20,000		NRAO
VLBA	I,R	0.3-90 cm	7.5 TB	3000	~3000	NRAO
GBT	S,I,R	0.3-90 cm				NRAO
DPOSS	I,C	4000-8500 Å	3 TB	3500 plates	1×10^9	Caltech
USNO-A2.0	I,C	4000-8500 Å	10 TB		5.2×10^8	USNO
MACHO	I,P,C	4500-7600 Å	7 TB	93,000	7×10^7	UPenn/LLNL
OGLE II	I,P,C	6800-8400 Å	1 TB	30,000	4×10^7	UPenn/Princeton
LOTIS	I	4000-8000 Å	12 TB	90,000	2×10^7	UPenn/LLNL
LONEOS	I,C	4000-8000 Å	2 TB	60,000	3×10^8	UPenn/LLNL
ADC	C	Various	18 GB	3500 catalogs		GSFC/ADC
NED	S,I,P,C,B	Various	155 GB	4.2×10^6	3.3×10^6	IPAC/NED
ADS	B		350 GB	>2.2×10^6 abstracts & references		SAO

A.2 Ongoing NVO Technology Development Efforts and Programs

A.2.1 NSF ITR, AST, DTF, and Other Programs

The National Science Foundation is supporting fundamental developments in computing and information technology that are key to the success of the National Virtual Observatory. For example, NSF's Information Technology Research program is aimed at fostering both new IT research and the application of that research to innovative projects. The FY2001 ITR program selected a major US initiative, "Building the Framework for the National Virtual Observatory," for funding over a five-year period. The ITR program also funded several smaller grants related to NVO technology and applications, and the FY2002 selection process is now underway. The Astronomy Division of the NSF is the point of contact for the NVO development efforts, and through the national observatories is now supporting the much-needed development of archival systems for ground-based data.

NSF's ITR program is also supporting the Grid Physics Network (GriPhyN) and International Virtual-Data Grid Laboratory (iVDGL) projects, efforts with similar technology goals and drivers as NVO but focused on high-energy physics data reduction and analysis.

In FY2000 NSF held a competition to assist in the establishment of a single new terascale computing system that would enable US researchers in all science and engineering disciplines to gain access to leading edge computing capabilities. On August 3, 2000 the National Science Board approved an award to the Pittsburgh Supercomputer Center, a partnership of the Carnegie Mellon University and the University of Pittsburgh. The 6 Teraflop peak computing system will be supplied by the Compaq Computer Corporation. In FY2001 NSF seeks to open a pathway to future computing, communications, and information environments by creating a very large-scale system that is part of the rapidly expanding computational Grid. NSF will establish an advanced, multi-site "distributed facility" connected by ultra high-speed networking that will lead to breakthroughs and enhance the capabilities of

U.S. researchers in all areas of computational, computer, and information science and engineering. The Distributed Terascale Facility will be fully coordinated with the resources and activities of the existing PACI partnerships.

The mission of the National Partnership for Advanced Computational Infrastructure (NPACI) is to advance science by creating a ubiquitous, continuous, and pervasive national computational infrastructure: the Grid. This infrastructure for the 21st century builds on dramatic advances in information technology to enable distributed research by interdisciplinary teams. In the NPACI vision, researchers collect data from experiments and digital libraries, analyze the data with models run on a computing grid, visualize and share those data over the Web, and publish the results for the scientific community in digital libraries.

Through overlap in personnel and newly established communications channels, the advances in information technology achieved through the NPACIs, the DTF, and various ITR projects will be brought to bear in creating and operating the NVO.

A.2.2 NASA AISRP, ADP, HPCC, and Other Programs

The NASA OSS Applied Information Systems Research Program (AISRP) maintains an awareness of emerging technologies applicable to space science disciplines, supports applied research in computer and information systems science and technology to enhance NASA OSS programs, stimulates application development where warranted, and provides for systems analysis and engineering required to transfer new technology into evolving OSS programs through NASA Research Announcements (NRAs). Specific areas of interest include high performance computing and networking, scientific data analysis and visualization, scientific data storage and management, and software technology including World-Wide Web tools. To date, the AISR program has been one of the strongest supporters of the technological developments leading to the NVO, with between a quarter and a third of all AISRP grants going to potential NVO-related technologies and tools.

To date there has been relatively little coordination among the various NASA-funded research efforts related to the NVO. Although AISRP grantees meet for an annual workshop to exchange ideas, their development efforts are not coordinated and they do not necessarily work toward a common programmatic goal. Similarly, through the HPCC NASA is developing the Information Power Grid, but thus far little has been done to find areas of common interest between this program and the NVO.

NASA also is supporting, at a small level, the development of a Virtual Solar Observatory to link NASA solar mission data sets and ground-based solar observations. The VSO will eventually replace the SDAC.

A.2.3 International Efforts

Astrophysical Virtual Observatory. The Astrophysical Virtual Observatory (AVO) project is a Phase-A, three-year study for the design and implementation of a virtual observatory for European astronomy. The AVO Proposal was submitted under the EC 5th Framework RTD scheme in February 2001. The European Commission has favorably reviewed the AVO proposal and is now proceeding with contract negotiations with the proposal team for a three year work program valued at approximately €4 million. The work program is planned to commence in 2002 and focus on the key areas of scientific requirements, interoperability and new technologies.

AVO Phase A will involve six partner organizations lead by the European Southern Observatory (ESO) in Garching near Munich. The other partner organizations are the ESA-operated Space Telescope European Coordinating Facility (ST-ECF) co-located with ESO, the ASTROGRID (UK) consortium, the CNRS-supported Centre de Données Astronomiques de Strasbourg (CDS) at the University Louis Pasteur in Strasbourg, the CNRS-supported TERAPIX astronomical data center at the Institut d'Astrophysique in Paris, and the Jodrell Bank Observatory of the Victoria University of Manchester. The AVO initiative was an outgrowth of interest and concern expressed within a network of European astronomical organizations collective known as OPTICON (Optical Infrared Coordination Network for Astronomy). This collection of 16 universities and research institutions was funded in 1999 by the EC under the Thematic Networks section of the 5th Framework Program. OPTICON has made it possible for groups of astronomers within Europe to have collaborative meetings in which alliances have been formed and new research initiatives have been proposed. One such initiative is the ASTRO-WISE RTD proposal that has also been favorably reviewed by the EC. Flowing from an OPTICON-sponsored workshop on survey systems in astronomy, ASTRO-WISE seeks to develop hardware and software systems for the production

and utilization of wide-field digital sky surveys that will be undertaken using the OmegaCAM instrument on the VST. ASTRO-WISE will be a key content provider for the AVO.

The AVO project is structured into three main development areas, each with an international working group. The *Science Working Group* will identify the scientific requirements for the AVO. Issues such as the quality and availability of a calibration pipeline and calibration data, the PSF or seeing characterization over the field of images, the limiting magnitude and limiting surface brightness versus exposure time and wavelength, photometric conditions, astrometry characterization, etc., need to be quantitatively addressed and verified against the existing archives. It is equally important to foresee the need of special scientific tools, such as algorithms to perform statistical analysis over a set of distributed databases without moving the data from where they reside. Since the "phase A" of AVO is planning to test different grid-type distributed infrastructure, it is essential that a realistic test bed of scientific applications be provided as soon as possible. The AVO science activity also includes the design of the operational or "phase B" of the project. In particular it will define the different ways in which the AVO can be accessed: either as an open on-line facility, similar but much more powerful than the current individual Archives, or as a "controlled" infrastructure that give access and support (e.g., with dedicated funds or specialized assistance) to large projects that are selected via a standard "Call for Proposals". The AVO SWG will address the above matters and others that the Group itself may indicate as relevant and it will monitor the pilot implementation of the AVO in terms of its actual and potential scientific use. The SWG will meet at least twice a year and will extensively use electronic means as a working environment.

The AVO *Interoperability Working Group* deals with issues concerning archive interoperability. The objective of this working group is to study and promote cost-effective tools and standards for improving access to data archives and information services, and for building links between distributed services, with a minimum workload on service providers. Discussions will be in partnership with the European radio community and will involve US and Canadian participants.

AstroGrid. AstroGrid is a £5M project aimed at building a data-grid for UK astronomy, which will form the UK contribution to a global Virtual Observatory. AstroGrid is also one of several **Grid** projects in the UK's E-science initiative. AstroGrid is funded via the UK's Particle Physics & Astronomy Research Council. The project was formally started on 2001 September 1, although development began early in 2000.

AstroVirtel. The ASTROVIRTEL Project, supported by the European Commission and managed by the ST-ECF on behalf of ESA and ESO, is aimed at enhancing the scientific return of the ST-ECF/ESO Archive. It offers the possibility to European Users to exploit it as a virtual telescope, retrieving and analyzing large quantities of data with the assistance of the Archive operators and personnel. The first call for proposals was issued in April 2000 and the second in April 2001 with a deadline for submission of June 15, 2001.

The goal of AstroVirtel is to foster archive-based science by sponsoring innovative research programs utilizing the data resources located at the ST-ECF and ESO. Already several significant scientific results have emerged, such as the discovery of the largest asteroid in the solar system.

EGSO: The European Grid of Solar Observations has been funded by the EC for three years (2002-2004) at a total level of 3 million Euros. The EGSO will build on a long-standing European interest in data management that reaches back to the original concept of a Whole Sun Catalog.

The Australian Virtual Observatory. The Australian Virtual Observatory will be an integral component of the international Virtual Observatory. The Australian contribution will be in four distinct areas:

- Providing Australian data to the rest of the world, to enable the best science to be done with, and highest international visibility to be achieved by, its existing cutting-edge facilities.

- Providing Australian researchers with a data grid to optimize access to data from overseas, facilitate data mining, enable the extraction of science from disparate data sets, and compare data with theoretical models.

- Developing software, techniques, standards, formats necessary for the establishment of the NVO.

- Upgrading Australian instrumentation where necessary to provide top-quality data necessary for the NVO.

Additional effort will be provided in establishing data standards, formats, compression techniques, and protocols, for transparent access of data across the NVO, setting up a national distributed database with adequate computing power as the prime server of leading edge Australian data (e.g. HIPASS, 2dFGRS, 2QZ, etc.) to the

world, and setting up high-bandwidth Internet links between the various data providers and users within Australia, and to the Australian user community and other data providers overseas, to provide a data grid so that the IVO can also inter-compare and manipulate different data sets.

Similar initiatives, or interest in such initiatives, are now being expressed Germany, India, Chile, and Japan.

A.3 Workshops and Conferences

In the short time since the release of the Decadal Survey, a number of conferences and workshops have been held to discuss the technical challenges and scientific benefits of the NVO. The first was *Virtual Observatories of the Future*, held at Caltech in June 2000. Bringing together ~150 astronomers and software developers, the opportunities and challenges of the virtual observatory were explored (see Brunner, R.J., Djorgovski, S.G., & Szalay, A.S. (editors) 2001, *Virtual Observatories of the Future,* Astronomical Society of the Pacific, Volume 225.) In August 2000, the first European conference on virtual observatories was held, in Garching, Germany. Entitled *Mining the Sky*, the conference focused on the science, tools, and technology needed to process and understand astronomical databases and datasets of unprecedented size (see Banday, A., *et al.* (editors) 2001, *Mining the Sky*, ESO Astrophysics Symposia, Berlin: Springer Verlag.)

The annual Astronomical Data Analysis Software and Systems (ADASS) conferences have featured NVO-enabling technologies for the past several years, culminating in ADASS XI (Victoria, BC, October 2001) where more than half of the three-day conference focused on the virtual observatory. Similarly, the SPIE conference (August 2001, San Diego) highlighted NVO science and technology in a session on astronomical data analysis. The science aspects of the virtual observatory were discussed at length in June 2001 at the Aspen Center for Physics conference *New Astrophysics and Cosmology with a Virtual Observatory*. The problems of managing large databases and understanding the statistical significance of cross-correlations between diverse catalogs were the subject of *Statistical Challenges in Modern Astronomy III* (Penn State, July 2001). In a special session at the January 2001 AAS meeting (San Diego) on the virtual observatory, attendance was standing-room-only. General exposure of the NVO concept to the European astronomical community occurred with the Joint European Nation Astronomy Meeting (JENAM) in September 2001 (Munich).

In the next two years a number of additional meetings are planned with a NVO focus. In June 2002, an international conference entitled *Toward an International Virtual Observatory* will be held in Garching, Germany. A special session on the NVO will also be held at the summer meeting of the AAS (Albuquerque, June 2002). The major SPIE meeting of the year, focused on astronomy and astronomical instrumentation, will have a dedicated two-day session on the NVO, as well as a complementary session on data analysis (Kona, HI, August 2002). ADASS XII (Baltimore, October 2002) promises to include many talks and posters concerning the initial development efforts on the NVO. Finally, a proposal has been submitted to the IAU for an international symposium on the NVO in conjunction with the next General Assembly (Sydney, August 2003).

Appendix B: Additional Examples of Scientific Case Studies

We present here more examples of scientific projects that will be greatly enhanced through the creation of a NVO. These case studies are a continuation of the examples given in Section 2.2 but are again, in no way complete.

B.1 Clusters of Galaxies as Cosmological Probes

Clusters of galaxies are the largest, gravitationally bound, masses in the Universe and as such, are sensitive probes of the underlying density and evolution of mass in the Universe. Therefore, the physical properties of clusters, as a function of redshift, can be used to place strong constraints on the cosmological parameters and models. For example, there is a strong dependence on the number density of clusters, with look-back time, on the mean density of the Universe and the normalization of the underlying density fluctuation spectrum. Such cluster observations are orthogonal to the CMB and Supernovae experiments and is thus one of the key components of the "Cosmic Concordance" model. To achieve these goals, we need to build massive catalogs of clusters that span a wide range in redshift and mass. These catalogs must possess well-understood selection functions (to allow the determination of the survey volume) and measure physically meaningful properties (like mass) for each cluster.

Outline of the Project:

- Identify clusters of galaxies from multi-wavelength pixel data using different physical signatures i.e.

    1. X-ray emission from a hot intracluster medium,

    2. Enhancement of optically red, elliptical galaxies,

    3. Scattering of CMB photons by the hot, intracluster medium (SZ effect), and

    4. The presence of a "bent" radio sources.

- Estimate the completeness and false-positive contamination of the catalogues. This can be achieved through adding fake clusters to the data and determining the success rate in detecting these systems. The false-positive contamination can be estimated using mock data sets.

- Compare the results of different selection techniques and understand any differences. Are there anomalous clusters e.g. luminous in X-rays, but undetected in optical and SZ. Check for potential contamination from non-cluster sources e.g. AGNs

- Quantifying the selection function for each methodology as functions of the clusters mass, density and redshift. This determines the cosmological volume sampled by each cluster catalog.

- Determine a robust redshift and mass, as well as other supplementary cluster properties, for each cluster.

- Compare results with simulations of the universe and determine confidence interval on cosmological parameters. Ideally, same cluster-finding algorithms would be run on mock optical, radio and X-ray maps of the Universe as a function of the simulation input parameters. This would be repeated many thousands of times to obtain an estimate of the cosmic variance.

NVO Functionality Required:

- Operate on large quantities of imaging data with user-defined algorithms and tools.

- Compare catalogs from different wavelengths accounting for measurement errors and upper limits.

- Construct simulated surveys to understand the selection functions; test user-defined tools on simulations.

- Generate predictions of observed sample properties based on various theories. Compare with observed samples.

B.2 Galaxies in N-dimensional Parameter Space

Galaxies span a huge range in physical parameter space, which include morphology, mass, luminosity, metallicity, star-formation rate, surface brightness etc. The "Holy Grail" of galaxy evolution studies is to understand the relative importance of Nature and Nurture on these physical properties and determine the main physical processes that are responsible for creating and evolving a galaxy. This grand endeavor has been greatly aided in recent years through large, multi-wavelength surveys of the Universe since the different spectral windows are sensitive to different physical phenomena and thus, we are able to determine the relative importance of the different physical effects. The Virtual Observatory will be essential for such science because it will allow astrophysicists to seamlessly correlate and compile a large set of galaxies that are detected across the electromagnetic spectrum. Furthermore, these multi-dimensional data will require new statistical tools to look for significant correlations and anomalies. Finally, these data will need to be statistically tested against theories of galaxy evolution.

Outline of the Project:

- Compile a panchromatic catalog of galaxies from radio through to X-rays. These data provide complementary information:

1. Radio measures the star-formation of galaxies in a dust-free way via the synchrotron radiation produced by supernovae of the young, massive stars.

2. Infrared data quantifies the global star-formation history of the galaxies by measuring the number of old stars in the galaxy.

3. X-ray surveys can probe the number of binary stars and the rate of supernovae, which in turn seed the formation of new stars.

4. Optical surveys provide the morphology and thus dynamics of the galaxies.

- Measure physical parameters for each galaxy, including morphology, star-formation rate, mass, metallicity etc. This will require significant processing of the data.

- Search for higher-order correlations in this multi-dimensional database. Seek the conditional properties of the physical parameter.

- Test against semi-analytical models of galaxy evolution.

NVO Functionality Required:

- Federation of relevant surveys including cross-identification of objects in multi-wavelength surveys and interchange/merging of metadata.

- Cluster analysis to identify statistically significant correlations and outliers. New and fast classification schemes for galaxies.

- Visualization of multi-dimensional datasets.

- Efficient creation of new semi-analytical simulations of the Universe with fast turn-around

- New statistical tools to effectively test the data against a suite of simulated galaxy catalogs.

B.3 Model of our Galaxy

The Galaxy is composed of a number of structural components i.e. halo, thin disk, thick disk, bulge, spiral arms. Each of these components can be characterized by a different population of stars (that possess a distribution of ages, masses, chemical compositions etc.) and non-stellar material like gas and dust. Taken together, these components provide a complete observational model for our Galaxy which can then be used to study the kinematics of the various Galactic components, searching for tidal debris tails, past merger events and previously undetected satellites. Such knowledge is critical for constraining theoretical models for the structure of the Galaxy, which are based on different formation processes. One can further constrain cosmological structure formation models like Cold Dark Matter. The construction of a detailed observation model of our Galaxy requires both a large sample of stars (each with information about their physical properties) and high-resolution maps of the dust and gas components.

Outline of the Project:

- Federate various optical and IR surveys of the sky to generate matched catalogs of stars.

- Use positions, magnitudes, and colors to construct three-dimensional stellar distributions. This will require using colors to derive luminosity classes and estimate extinction.

- Quantify dust distribution and obscuration using FIR, H I, and CO images.

- Identify bulk flows and sites of star formation using IR and radio images.

- Use proper motion surveys (and radial velocity information, when available) to deduce motions of subsets of stars.

- Use multi-epoch imaging to find variables. Use these to provide a distance check.

NVO Functionality Required:

- Statistical tools for the federation of multi-wavelength and multi-epoch catalog data.

- Facility the correlation pixelized map data with point-like catalog data.

- Visualization tools for handling large multi-dimensional datasets of both images and catalogs of points.

- Automated knowledge discovery tools to aid the scientists to find and quantify correlations within this observation model.

B.4 Asteroids, Comets, and Kuiper Belt Objects

The solar system possesses hundreds of thousands of detectable objects, clustered towards the ecliptic plane, but found in all regions of the sky. The bulk of these are fragments of protoplanetary bodies, which formed between Mars and Jupiter, whose composition and physical properties provide insights in to planetary formation processes. Others formed at far greater heliocentric distances, representing primordial compositions and processes that may be replicated around other stars. To advance our understanding of the solar, astronomers require a compositional catalog of small solar system bodies in order to test theories of primordial compositional gradients in the early solar system and how it was modified through heating (from solar T-Tauri or radioactive nuclide decay) and planetary formation processes. In addition evidence of compositional scattering may preserve an indirect record of early planetary migration.

Outline of the Project:

- Identify Optical/near-IR/Thermal survey sources with solar system asteroids/comets/KBOs with known orbits.

- Create pan-spectral catalogs correlated with mineralogical spectral models and albedo/diameter models.

- Construct a compositional atlas of the solar system. Compare with predictions from cosmochemical and dynamical models of the early solar system.

NVO Functionality Required:

- Cross-identification of objects across multiple surveys of different wavelengths on the basis of predicted positions for specific times. High precision temporal information must be preserved and recorded.

- Merge selected information from different NVO components that are correlated with a given set of objects.

B.5 Understanding the origin of the solar wind

The solar wind is a high-speed stream of energetic and electrically charged particles that is continually ejected from the Sun. This wind shapes the interplanetary magnetic field, interacts with all of the solar system components, and affects the terrestrial environment through interactions with the ionosphere. These interactions range from the harmless yet beautiful auroral displays to severe magnetic storms that disrupt modern technology such as telecommunication and power systems. The generation mechanism of the solar wind is not fully understood. Progress in this area would lead to further understanding of a basic solar process, and could provide a predictive tool for forecasting space weather. It would also contribute to astrophysics, since stellar winds play a fundamental role in distributing mass and magnetic fields into the interstellar medium, affecting star formation and the galactic magnetic field. To achieve this goal, we need the ability to combine several disparate solar data sets ranging from images to spectra to particle counts, perform complex pattern recognition searches, and create and visualize multi-dimensional time series.

Outline of the Project:

- Collate and collect solar data:
    - Helium 1083-nm or x-ray images to locate the coronal holes from which the wind is emitted,
    - High-resolution spectra in several wavelengths to measure properties in the solar atmosphere,
    - Doppler velocity images to map surface flow patterns,
    - Local helioseismology data to probe the solar interior below the coronal hole,
    - Coronal density images to map magnetic field patterns and coronal mass ejections (CMEs),
    - Images of the surface magnetic field,
    - Particle counts near the earth to measure the wind velocity,
    - Interplanetary magnetic field measurements near the earth.
- Apply pattern recognition to identify and locate coronal holes boundaries, magnetically active regions, and flow patterns. Can a class of patterns spanning these (and other) data sets be correlated with solar wind characteristics?
- Track solar properties above and below coronal holes using helioseismology for subsurface measurements and spectra for atmospheric properties.
- Follow wind properties from solar surface to the earth using the coronal images, particle counts, and interplanetary magnetic field measurements.
- Create and visualize correlated time series of relevant quantities and images. Perform cross-correlation and relative phase analyses to investigate underlying mechanisms.

NVO Functionality Required:
- Federation of multi-wavelength and multi-observable solar datasets.
- Sophisticated and flexible pattern recognition tools.
- Tools to register and geometrically transform data between qualitatively different coordinate systems and time bases.
- Visualization tools for handling large multi-dimensional and time-dependent datasets of images, spectra, and 1-d time series
- Spectral analysis, cross-correlation, and other time series analysis tools

B.6 Gamma-Ray Bursters

Once the satellite Swift is launched in the fall of 2003, the rate of Gamma-Ray Burst (GRBs) will increase dramatically from ~1 per month (localized to tens of arcminutes within hours) to ~1 per day (localized to arcminutes within seconds, and arcseconds within minutes). This will force a dramatic change in the way the follow-up of these GRBs is carried out: One approach will be to target special types of bursts e.g. very high redshift bursts and highly extinguished bursts. The downside of this approach is that new and interesting types of bursts (such as GRB 980425/SN 1998bw) are likely to be missed. Consequently, the other approach will be to chase every accessible burst with dedicated, automated telescopes. Currently, such telescopes are being built on half-meter scales (e.g., ROTSE III, Super Lotis) and, when they are not chasing GRBs, they will be performing multi-wavelength surveys of the sky. Thus, the search for GRBs, and the telescopes that perform this task, lead for a natural pairing of GRB science and NVO science is natural i.e. the data can be automatically and immediately made accessible through the NVO architecture.

A second way in which NVO science relates to, and indeed directly affects, GRB science is in the discovery of orphan afterglows. We now know that the bursts are highly collimated: For every burst that is pointed toward us (which we see), there are ~500 bursts that are pointed away from us (which we do not see). However, these jets expand laterally with time. Consequently, there are many bursts that although we cannot see them in gamma rays,

we could see them at optical wavelengths hours to days after the burst if we only knew where to look. These orphan afterglows will only be detected in large surveys of the sky, particularly if they are repeated periodically.

B.7 Cosmic Web and Large-Scale Structure in the Universe

The origin of the large-scale structures in the Universe is one of the last remaining puzzles of our cosmological model; how did the complex structures we see in the local universe form from smooth initial plasma. It is commonly believed that these massive structures we initially seeded by quantum fluctuations during the Inflation era of the Universe. To test this and other theories of large-scale structure formation, astronomers have embarked upon massive new surveys of the local cosmos collecting redshifts for millions of galaxies e.g. the SDSS. These surveys will produce a complex picture of the distribution of galaxies that will be full of non-linear structures. These maps must be tested against simulations of the universe as to determine which of our models is successful. To date such comparison of simulations and data has revolved around simple, first-order statistical tests like the 2-point correlation function or the mean and skewness of counts-in-cells. These statistical measures hide much of the complex structure in these maps and in the simulations. To move beyond this present situation, cosmologists will need new, more sensitive analysis tools as well as quick and efficient access to massive surveys of the sky and a suite of simulations.

Appendix C: Designing the NVO

The design and implementation of the NVO would benefit from the great recent and ongoing advances in information technology and computer science, but still involves significant technical challenges. These include both the incorporation of existing and future data archiving efforts in astronomy as well as the development of new capabilities and structures. Major technical components to the NVO include archives, metadata standards, a data access layer, query and computing services, and data mining applications. Development of these capabilities will require close interaction and collaboration with the information technology, applied computer science and statistics communities. Paramount to this entire effort is the need to maintain focus on the technical requirements that are needed to satisfy the scientific drivers.

In order for the NVO to become a successful tool for the astrophysical community, it must be scientifically viable. This implies that the first task when designing the NVO is to explore the technical requirements that are necessary to accomplish science within the NVO framework. These requirements that cover everything from efficiently archiving data to federating highly distributed archives to processing and analyzing large quantities of complex data, can be distilled into three key areas: **data persistence**, **data communication**, and **data processing**.

C.1 Data Persistence Requirements

In order to be effective and attractive potential contributors, the NVO must not impose unnecessary bureaucratic restrictions on specific persistence mechanisms used by data providers. This is a natural consequence of the fact that data providers, be they NASA data centers, survey archives, or individual astronomers, may be under pre-existing restrictions to use specific hardware or software. Furthermore, these data providers may also have needs outside the scope of the NVO to utilize specific practices or techniques particular to a given wavelength domain, for example, the differing uses of the HTM (see http://www.sdss.jhu.edu/htm/) or the CMB Healpix (see http://www.eso.org/science/healpix/) methods.

Instead, the NVO must specify requirements that dictate how a data provider can participate in the NVO. These requirements can be distilled into providing standard interfaces to first access the metadata for a given data product, and second to access the actual data. If a data provider also wishes to make available specific services, the mechanism for finding and accessing these services also must be articulated using a standard format. This represents a natural extension of the successful standardization efforts in the past such as the FITS image format, or the ADS bibcode.

C.1.1 Metadata Requirements

Metadata (literally, "data about data") is structured information describing some element of the NVO. Metadata will be required to describe archives, the services provided by those archives, the data collections available from an archive, the structure and semantics (meaning) of individual data collections, and the structure and semantics of individual datasets within a collection. Typical astronomical datasets are data objects such as catalogs, images, or spectra. As an example, the semantic metadata for a typical astronomical image is the logical content of the FITS header of the image.

Metadata describing astronomical data are essential to enable data discovery and data interoperability. Metadata describing archives and services are necessary to allow the components of NVO to interoperate in an automated fashion. Metadata standards are desirable to make these problems more tractable. In practice there are limits to what can be done to standardize dataset specific metadata, but mediation techniques such as those being developed by the digital library community (see http://www.diglib.org/) provide ways to combine metadata dialects developed by different communities for similar types of data. Current projects such as Astrobrowse and ISAIA (Interoperable Systems for Archival Information Access) represent initial efforts within the astronomical community to establish metadata standards.

C.1.2 Data Access Requirements

A successful NVO should place no requirements on data archives other than that they implement the formal NVO standard interfaces for communicating both metadata and data as well as publicly available services. This is required so that different NVO participants, be they other data providers or tools, can find and interact with a specific data provider. In the simplest cases, interfacing an archive to NVO will be little more than a matter of installing data access software and modifying a few configuration files to reflect the data holdings and access permissions of the local archive, much as one would install a WWW server. More sophisticated installations may provide expanded support for metadata access and server-side functions.

The data access functionality will provide a uniform interface to all data, metadata, and compute services within NVO. At its lowest level, this access layer is only a standard protocol defining how the software components of the NVO talk to each other. Reference grade software implementing the protocol will also be provided, which can either be used directly or taken as the basis for further development by the community. This software will include server-side software used to interface archives and compute services to the NVO, and client-side applications programming interfaces (APIs), which can be used to write NVO-aware distributed data mining applications. Since it is fundamentally a protocol, multiple APIs will be possible, e.g., to support legacy software or multiple language environments.

The key aspect for data access is the existence of a uniform interface to all data and services within NVO. User applications use the standard interface to access NVO data and services, and archives and compute services within the NVO use it internally to access data or services in other archives, potentially generating a cascade of such references. NVO is thus an inherently hierarchical, distributed system, which nonetheless has a simple structure since all components share the same interface. In addition to such location transparency, this technique will provide storage transparency, hiding the details of how data are stored within an archive. Finally, the data access protocol will define standard data models (at the protocol level) for astronomical data objects such as images and spectra. Archive maintainers will provide server-side modules to perform data model translation when data objects are accessed, allowing applications to process remote data regardless of its source or how it is stored within a particular archive.

Often a client program using the data access APIs will not need an entire dataset, but only a portion. Server-side functions will permit subsetting, filtering, and data model translation of individual datasets. In some cases user defined functions may be downloaded and applied to the data to compute the result returned to the remote client. This is critical to reduce network loading and distribute computation.

Since both metadata and actual datasets can be retrieved from a remote archive, dataset replication becomes possible, allowing a local data cache to be maintained. This is critical to optimizing data access throughout the NVO, and will be necessary to even attempt many large-scale statistical studies and correlations. Dataset replication also makes it possible to replicate entire data collections, and to migrate data archives forward in time. Metadata replication and ingest makes it possible for a central site to automatically index entire remote archives.

C.1.3 Archive Creation Requirements

Experience over the past decade has shown that astronomical archives are complex and diverse, never stop growing, and are best maintained by those close to the data who know it well. In practice this has meant that most data are put online either by individual large survey projects, e.g., the 2-Micron All-Sky Survey (2MASS) or the Sloan Digital Sky Survey (SDSS), or by discipline-specific archive centers, which serve a given community. To address the need to move to large scale archiving of ground-based astronomical data, archiving facilities will need to be established at both the national centers (NOAO, NRAO, NSO, NAIC) and the major private and university-operated facilities. The major national data centers for ground- and space-based data will comprise the principal nodes of the distributed NVO data system in the U.S.

Any consideration of the science to be performed by the NVO, or the technical issues involved in implementing the NVO, must start with the data. Although the data from most NASA missions have been routinely archived for over a decade, relatively little data from ground based telescopes is currently available online, other than for a few major surveys. With modern wide-field and multispectra instruments on ground-based telescopes producing ever-larger quantities of data, and with ground-based survey projects becoming almost as common as classical observing, there is an acute need to archive and publish high quality datasets from ground-based instruments and surveys. The science promised by the NVO will not be possible unless the NVO succeeds in creating true, panchromatic images and catalogs, seamlessly integrating data from both ground- and space-based archives, and thereby enabling exploration of astrophysical phenomena over most of the electromagnetic spectrum. In order to defray the cost of joining the NVO, "archive templates" may be developed and utilized to simplify the potentially daunting task of archiving and serving data within the NVO.

C.2 Data Communication Requirements

Since astrophysical data is highly distributed, especially given the growing global data availability, a successful NVO places strong requirements on the transferring of data between data providers, service providers, and the end-user. First, in order to move the large quantities of data between the various data providers and compute facilities, wide bandwidth is a must; however, end-users should be able to access the end-results of their queries through their existing Internet connections. Second, standard protocols for transferring data between interested parties must be developed and deployed. Finally, the entire system must employ management functions that provide access control, account for varying network accessibility, and allow for the growth of the entire system.

C.2.1 High Speed Communication Requirements

The large dataset size and geographic distribution of users and resources also presents major challenges in connectivity. Next generation networking providing cross-continental bandwidths of 100 MB/sec is now available and currently underutilized, but this situation will change rapidly. It will be essential for the major NVO data centers to be interconnected with very high speed networks (see http://www.internet2.edu/), and to utilize intelligent server-side software agents in order to make the most efficient use of the network when interacting with end-users.

C.2.2 Communication Protocol Requirements

To meet this wide range of requirements, the NVO needs a distributed system architecture that provides uniform and efficient access to data and services irrespective of location or implementation. Data archives are assumed to already exist and will vary considerably in implementation and access policy. Metadata standards will be devised to provide a well-defined means to describe archives, data collections, and services. A data access layer will provide a single uniform interface to all data and services, and will be used both to link archives and services within the framework of NVO, and to allow user applications to access NVO resources. Query and compute services will provide the tools for information discovery and large-scale correlation and analysis of disparate datasets. Data mining applications, running on a user workstation at their home institution, as applets within a WWW browser, or at a major NVO data center, will provide the main user interface to enable science with the NVO.

Data archives store datasets (e.g., catalogs, images, and spectra) organized into logically related data collections, as well as metadata describing the archive and its data holdings. Access is provided in various ways

such as via a structured WWW interface, via a standard file-oriented interface such as FTP, or via other access protocols that may vary from archive to archive.

C.2.3 Communication Management Requirements

While data will be widely distributed, the large studies at the cutting edge of the science to be enabled by the NVO will need massive computational resources and fast local access to data. While sophisticated metadata standards and access protocols will be required to link together distributed archives and network services, the effort required to interface a small archive to the NVO must be minimized to encourage publication of new data collections by the community. While data collections and compute services will be widely distributed, users will need a straightforward interface to the system which makes the location and storage representation of data and services as transparent as possible.

While the data access and metadata standards will allow the NVO to link archives and access data, query and compute services will be required to support information discovery and provide the statistical correlation and image analysis capabilities required for data mining.

The vast amounts of information and available services will swamp the new user. Currently, an astronomer generally must spend a non-negligible amount of time learning how to use an archive or a new analysis tool. Multiply this situation by the number of available archives and tools, and the complexity of working in the era of large database becomes very daunting. In order to improve the efficacy of working within the NVO framework, data and service registries will be created that allow standard mechanisms for finding and interacting with different datasets or services. Since all of the necessary information to facilitate this scenario is available in the metadata and metaservice interfaces, advanced tools can be developed to generate complex queries and analysis by plugging together data providers and processors. A good model for this is the Web Services paradigm, where services are registered using UDDI (see http://www.uddi.org/) and are described using WSDL (see http://www.w3.org/TR/wsdl). Web services can be connected together into a single transaction using WSFL (see http://www-4.ibm.com/software/solutions/webservices/pdf/WSFL.pdf).

C.3 Data Processing Requirements

Given archival quality data from all branches of astronomy, physically distributed at several tens of major archive centers and a much larger number of ancillary datasets together with a distributed community of thousands of scientific users, one can define what new software and services will be required to implement the NVO. Analysis of the data will be complex, due to the heterogeneous nature of datasets from the different branches of astronomy and due to the use of increasingly complex data structures (for example, extended FITS, HDF, or VOTable standard data formats) to accommodate the increasing levels of sophistication of modern astronomical instrumentation.

The sheer scale of the problem is daunting, with catalog sizes approaching the Terabyte range and the total data volume in the Petabyte range. However, an even more serious challenge comes from the complexity of these datasets, with tens or hundreds of attributes being measured for each of ten million or more objects. This is a crucial new aspect to the data mining issue, and multivariate correlation of such large catalogs is a massive computational problem. If pixel level analysis of candidate objects is required the computational problem can be even more massive. It is important to recognize that current brute-force analysis techniques do not scale to problems of this size! Multidisciplinary research in areas such as metadata representation and handling, large scale statistical analysis and correlations, and distributed parallel computational techniques will be required to address the unprecedented data access and computational problems faced by the NVO.

C.3.1 Data Federation

While most archives will provide basic query services for the data collections they support, large-scale data mining does not become possible until multiple catalogs are combined (correlated) to search for objects matching some statistical signature. The larger NVO data centers will provide the data and computational resources required to support such large-scale source cross-identification. While a specific query may result in subqueries to remote archives, extensive use of dataset replication and caching will be employed to optimize queries for commonly

accessed catalogs or archives. Sophisticated metadata mediation techniques will be required to combine the results from different catalogs.

Fundamentally, the problem of source cross-identification seems rather trivial, why not do one big cross-correlation between all of the available data providers? Simply put, the methodology, if not the very results, varies according to the application; and, we end up, therefore, with a dynamic result. Whether it is due to different resolutions, changing calibrations, or different physics, users will want to fine-control identification algorithms that federate multi-wavelength data. With the forthcoming multi-epoch surveys, this situation will only become even more acute.

C.3.2 Data Reprocessing

In some cases pixel-level analysis of the original processed data, using an algorithmic function downloaded by the user, may be required to compute new object parameters to refine a parametric search (in effect this operation is dynamically adding columns to an existing catalog, an extremely powerful technique). Since with NVO candidate object lists may contain several hundred million objects, this is a massively parallel problem such as might require a Terascale supercomputer to address. Even in the case of large scale statistical studies, distributed computing techniques and fast networks may allow the user to work from their home institution, but some form of peer reviewed time allocation may be required to allocate the necessary computational and storage resources. For some larger studies users may need to visit a NVO data center in order to have efficient access to personnel as well as data, software, and computational resources.

C.3.3 Data-Mining

The field of data mining, including visualization and statistical analysis of large multivariate datasets, is still in its infancy. This will be an area of active research for many years to come. Most current astronomical data analysis software will need to be upgraded to become "NVO-aware", able to be used equally well on both local and remote data. New applications will be developed as part of ongoing research into data mining techniques. While NVO should provide the interfaces and toolkits required to support this development, as well as some initial data mining applications from the major NVO centers, the open-ended nature of the problem suggests the need for a multidisciplinary data mining research grants program once NVO becomes operational.

Effective and powerful data visualization will be an essential part of the NVO. The technical challenge of visualizing large datasets, either in the image or even catalog domains, is an active area of research. A new area is the visualization of streaming data, which could easily exceed a Terabyte. At the other end of the scale are the difficulties that are present when trying to visualize the catalog domain where each data vector may have a hundred or more attributes. At some stage, the data analysis algorithms and the visualization techniques may need to be merged to better facilitate data exploration. This will improve the feedback between the end-user and the NVO and facilitate the utilization of the underlying NVO facilities for advanced data exploration processing, while trial-and-error exploration is performed locally.

Interpretation of data recovered through data mining is of little use without analysis. At present, analysis tools match the capabilities of the data mining. New data mining capabilities must be accompanied by new analysis capabilities in order to recover the science. These tools must be able to work with the data structures delivered by data queries to the NVO, be extensible (e.g., have an appropriate API), and include the core analysis functions such as visualization, fitting, etc. (This is hard, e.g., IRAF, AIPS, AIPS++, CFITSIO, CIAO, STSDAS, etc., have all been produced and form an existing base of analysis and processing code. NVO should build on these by defining standards for data structures, and allow retrofitting of data models where possible).

C.3.4 The Grid Paradigm

As we move into the twenty-first century, a new model for performing computational intensive scientific research is beginning to emerge. In this model, data storage facilities and data compute services are linked together into a grid. One of the leading efforts in this area is the Grid Physics Network, or Griphyn (see http://www.griphyn.org/). Key to the grid concept is the concept of virtual data, where the physical location of the data (or even its very existence) is irrelevant to the end-user. The International Virtual Data grid laboratory is conducting research in this area (see http://www.ivdgl.org/). One final facility, the Distributed Terascale Facility, or

DTF, see http://www.teragrid.org) is an NSF funded facility to construct a grid composed of distributed teraflops compute services and Petabyte storage resources.

C.4 Design Summary

In all areas-storage technology, information management, data handling, distributed and parallel computing, high speed networking, data visualization, data mining-NVO will push the limits of current technology. Partnerships with academia and industry will be necessary to research and develop the information systems technology necessary to implement NVO. Collaborations with other branches of science and with the national supercomputer centers will be required to develop standards for metadata handling, data handling and distributed computing. Data mining is an inherently multidisciplinary problem, which will require the partnership of astronomers, computer scientists, mathematicians, and software professionals to address

A next generation, high speed national research Internet is already in place, but is underutilized at present due to the lack of credible academic applications designed to make use of high performance networking. NVO would be a prime example of a creative new way to use wide area high performance networking for academic research.

Initial and continuing design issues for the NVO should be based on a sound design review process. NVO design should be considered as a dynamic and evolving structure capable of incorporating the latest developments in information technology and responding to the evolving scientific needs of the astronomical community.

Appendix D: Lessons Learned

The international astronomy community has embarked on related projects in the past with similar or complementary goals to the NVO. Notable examples are the Astrophysics Data System (ADS) (sponsored by NASA in the late 1980s and early 1990s), the European Space Information System (sponsored by ESA in the same time period), Interoperable Systems for Archival Information Access (ISAIA) (a small NASA AISRP-funded exploratory project, 1999-2000), and Astrobrowse, a data location service developed by collaborating NASA and European data centers within existing funding guidelines (1998-2000). NASA has also supported other data integration services such as the NASA Master Directory for space science missions, and has been implementing EOSDIS over the last decade to provide data and information management for the depth and diversity of earth science data sets. In 1998-2000, NASA's Space Science Data Systems Technical Working Group actively explored the technologies and management strategies needed to provide integrated access to space physics, planetary, and astrophysical data.

There have been numerous studies carried out within the NASA community, in Europe, by the National Academy of Sciences, and elsewhere to assess the successes and failures of these projects. What have we learned from these projects, studies, and committees? In particular, what is different about NVO? Why should it succeed when other data integration initiatives have not, or have promised more than they could deliver?

First, we have learned that timing is critical. ADS and ESIS were well-intentioned initiatives. Indeed, in reviewing their project goals, much of the text would translate into NVO goals verbatim-*the concepts driving ADS and ESIS remain valid today.* However, both projects, at least in their original scope, were eventually cancelled.

Why? ADS and ESIS both required technology ("middleware") that was not generally available circa 1990. Both had to rely on new, and as it turned out-proprietary-software for the communications layer between data centers/services and the end-user. Both projects pre-dated the full emergence of the World Wide Web. The project leaders were visionary, but lacked the tools needed to build the necessary infrastructure. In the decade since these projects were conceived, the WWW has matured, and associated middleware technologies (HTTP, HTML, XML, SOAP) have developed to the extent that little custom software needs to be developed in order to support distributed queries to multiple data centers and services.

A corollary to the timing issue is that because ADS and ESIS could not rely on industry standard middleware, they were required to invent their own. Their inventions required data providers to adhere to *internal* database standards. Rather than define a metadata standard that transcended any site-specific implementation, they mandated adherence to an external standard for internal data representations. This was unacceptable to many projects, for

which the cost of translation of internal databases to a new externally defined standard would have been prohibitive (and *not* paid for by the projects!).

ADS, which originated from NASA-sponsored community-based discussions in 1987-1988, also suffered from an overly centralized development approach. The concept of shared development and consensus on interoperability standards was inadvertently thwarted by an overly top-down management structure. Interoperability standards are not created by management fiat. The astronomy community's FITS standard is a case in point. The process for amending the FITS standard is slow and deliberate. While frustrating to some, this process allows time for community feedback and consensus building. It is probably safe to say that no other professional community has reached the level of data interchange standards (both syntax and semantics) that we have reached in astronomy.

ADS succeeded in a particular subset of its original scope, in providing standards for indexing and cross-referencing the astronomical literature. The ADS bibliographic service is now without peer, and is the most heavily used astronomical information service worldwide. It succeeded in this area for several reasons:

- Publishers and producers of the astronomical literature had a strong will and incentive to link their resources.
- A simple but clever encoding scheme-determinate, easily encoded, easily decoded (and human readable), was developed and remains the mainstay of the system (the "bibcode").
- This encoding scheme did not have to replace, or displace, something else. It was new and welcomed by information providers.

The lessons-learned from ADS for the NVO are clear: build upon the desires of astronomy data centers to participate in the NVO framework, keep the encoding standards simple, build all NVO middleware components from open, freely available software and tools, and do not force data providers to change internal systems. The potential barrier for participation in the NVO must be kept low, and the benefits kept high.

In addition to problems similar to those faced by ADS, ESIS also suffered from setting its scope too broadly. ESIS aimed to provide integrated access to a broad range of astronomy and space physics data. Unfortunately, the level of commonality between astronomy and space physics experimentation and exploration is modest at best, and more commonly minimal to non-existent. This problem was rediscovered ten years later in the ISAIA exploratory project. The diversity in utilization modes and in data paradigms creates substantial barriers to system design, common standards, and integration of existing facilities. Yes, there are interesting and compelling scientific investigations that require the integration of astronomy and space physics data sets. These are a small minority, however, of the scientific questions being asked, and this is not because of the difficulty in integrating the data sets. Astronomy is primarily asking questions of the remote universe, beyond the solar system, and matters of interest to space physicists are akin to the interests of meteorologists (from the astronomer's perspective). Astronomers want to see beyond the local "weather", be it the earth's atmosphere or the solar wind, and although both phenomena are intriguing in and of themselves, they are not the focus, nor common science goal, of the astronomer. Some notable exceptions exist, such as the astrophysics of comets, the understanding of whose behavior requires integration of data from remote sensing instruments on the ground and in space, direct imaging and spectroscopy, in situ measurements of the solar wind, and an integration of astronomical and space physics data sets. It is not cost-effective, however, for rather unique science problems to drive the design of a more general system. It is also difficult to implement a system to integrate data sets that have few if any common characteristics.

Both ADS and ESIS also suffered from sociological problems. Both were seen in the community as technology projects, motivated, driven, and controlled by computer scientists and astronomers interested primarily in what could be done without adequate regard for what needed to be done. At this point in time, NVO has some of the same problems. While state-of-the-art information technology is essential to the success of NVO, and while the IT community is a major partner with the astronomy community in NVO, the onus is on the astronomers who understand the scientific potential of NVO to clarify the goals in compelling scientific terms. Some, perhaps the majority, of the benefits of NVO may well be unpredictable. Who would have seen that the end result of ADS would be the bibliographic/abstract service that the worldwide astronomy community now depends on, now uses so widely as to reduce the time required for reference searching and finding new research results by factors of 10–100? Who would have known that NASA's modest investment in the Extragalactic Database project at Caltech would have provided a tool to research astronomers that would increase their scientific productivity by an order of magnitude?

The Astrobrowse project was the initiative of a few astronomer/technologists who realized that the WWW would enable elements of the original vision of ADS and ESIS to be implemented without having to develop the

underlying technology from scratch. With no specific funding, motivated staff at HEASARC, STScI, and CDS developed a data location service that demonstrated the power of "one-stop shopping". One query – "tell me who has data on NGC 1068?" – can result in immediate access to dozens of potential gold mines.

The ISAIA pilot project cannot be said to have failed, as it was never funded at a level sufficient to implement software. The ISAIA team learned two important lessons that are now being applied to the design of the NVO. First, as mentioned earlier, it is neither cost-effective nor scientifically very valuable to set scope too broadly. NVO focuses on astronomy (although its components may well be usable in other disciplines, just as NVO borrows components already developed in other contexts). Second, ISAIA highlighted the importance of metadata standards and explored frameworks for definition, distribution, and maintenance of those standards. This experience will have a direct bearing on the development of the NVO/AVO/AstroGrid metadata standards.

The Earth Observing System Data and Information System (EOSDIS), later the Earth Observing System Data and Information System and Services, is an example of a project in another discipline that sought to unite diverse data sets and a very diverse user community with a single system. It has also had a difficult history, is not regarded as successful in meeting many of its goals, and has been subject to high level of scrutiny by NASA, Congress, and the earth sciences community. EOSDIS had the added complication of having to serve operational needs of the Earth Observing System, providing data pipeline processing to a variety of very high data rate satellite experiments, while simultaneously joining archive and research centers and supporting a large research community. Possibly because of the operational considerations, ESODIS, although conceived as a distributed system and with functions distributed over multiple sites, was designed in a very centralized manner. Even after mid-course corrections, and after recommendations to consider the Internet paradigm, it remained essentially a centralized system duplicated at multiple sites. The operational component (the EOSDIS Core System) dominated the program and the distributed users felt disenfranchised. It was also centrally developed by a single major contractor, who solicited and merged the diverse requirements, and then built the system. Thus it not only suffered from the problems of diverse requirements, distributed functionality, and operational requirements, but also from the canonical problem of other major government IT procurements, wherein technology changes on a timescale faster than the procurement and development lifecycles.

The NVO development paradigm is very different from that of EOSDIS, and indeed from the previous efforts like ADAS and ESIS. NVO starts out as a "glue" to tie together existing, functioning data centers. These data centers are responsible for their own operational components (pipelines, etc.) and are free to maintain their own internal database formats and user interfaces. As part of NVO, they are voluntarily collaborating, adopting common interchange standards, communications protocols, etc. ("middleware"). They are sharing in defining the requirements for and designing the overall system and in developing new tools to make their data accessible to the overall system, and to provide new scientific capabilities from the system as a whole. Despite the variety of types of datasets and catalogs in the various sub-disciplines of astronomy, all the centers support a common user community. Unlike some of the earlier efforts, active researchers in the astronomy community, not IT specialists, conceived the NVO. It received strong endorsement by the National Academy of Sciences, as the highest priority "small" astronomy and astrophysics project of the next decade. After the previous "fits and starts" of large-scale distributed data systems, NVO appears to be an implementation whose time has finally come.

Appendix E: A Plausible Budget for the NVO Development

The following table shows a plausible budget for the NVO initiative. This budget does not allocate specific portions of the program to one agency or another, with two exceptions: (1) the current NSF-funded NVO framework development project is included, showing its relationship to a more general program, and (2) estimates are included for the costs of bringing the archives of the NSF-funded national observatories on-line and NVO-capable, and supporting their ongoing maintenance and operations. In the latter case, the budget represents NVO-incremental costs to data management support and services and is not intended to represent all data management expenses of the national facilities.

The assumed labor and inflation rates are listed at the bottom of the table. Inflation-adjusted figures are rounded to the nearest $1,000.

(1) The NSF ITR project funding is shown in total. The management, development, and education/outreach components of this project are not broken out and included in these categories elsewhere.

(2) The operations budget includes explicit costs for high-speed network connections. These might also be provided as part of a national network infrastructure. Hardware and software costs are only those needed to augment existing archive and information services to support the NVO.

(3) The Survey and Data Access Grants Program is intended to assist groups providing survey data with sufficient funds to assure the on-line availability of these data to the NVO. These funds would be allocated through a competitive, peer-reviewed process.

(4) The Research Grants budget is intended to foster community-based research using NVO resources. Its management and budget is patterned on programs such as the HST and Chandra GO/AR programs. The steady-state funding level of $1.5M/year is envisioned to support 20-30 research projects.

(5) The Fellowship Programs budget provides funding for fellowships at the post-doctoral, doctoral, and undergraduate level. The steady-state funding level of $1.0M year is based on supporting 10-15 FTEs, though the exact number would depend on the mix of these three levels.

(5) The Education and Public Outreach budget is substantial, representing over 6% of the total initiative. We believe this is appropriate given the great potential for engaging the public and enhancing educational opportunities that the NVO provides. The EPO program starts early and ramps up, commensurate with the goal of fully integrating these activities into the NVO framework.

(6) Management and Administrative costs are intended to be modest.

The ten-year integral cost estimate is $76M, or $90M when inflation adjusted at a rate of 3.5% per annum. This is not necessarily all "new money," as some elements of the NVO may well come from existing technology development programs.

NVO Budget Prototype (March 2002)

Item	2002 FTE	2002 $	2003 FTE	2003 $	2004 FTE	2004 $	2005 FTE	2005 $	2006 FTE	2006 $	2007 FTE	2007 $	2008 FTE	2008 $	2009 FTE	2009 $	2010 FTE	2010 $	2011 FTE	2011 $	Totals
Bring ground-based archives on-line																					
o Development	6	$900,000	6	$900,000	3	$450,000	1	$150,000													$2,400,000
o Maintenance					1	$150,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	$2,250,000
o Computer hardware		$200,000		$150,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000	$750,000
o Computer software		$20,000		$20,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000		$10,000	$120,000
o Operations			1	$150,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	2	$300,000	$2,550,000
Subtotal, ground-based archives		$1,120,000		$1,220,000		$960,000		$810,000		$660,000		$660,000		$660,000		$660,000		$660,000		$660,000	$8,070,000
NVO Development																					
o NSF ITR project	13	$2,000,000	18	$2,750,000	12	$1,840,000	12	$1,750,000	11	$1,660,000											$10,000,000
o Follow-on development					4	$600,000	4	$600,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	6	$900,000	$6,600,000
Subtotal, development		$2,000,000		$2,750,000		$2,440,000		$2,350,000		$2,560,000		$900,000		$900,000		$900,000		$900,000		$900,000	$16,600,000
Software and systems maintenance							4	$600,000	4	$600,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	8	$1,200,000	$7,200,000
Operations																					
o Support staff					2	$180,000	4	$360,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	6	$540,000	$3,780,000
o Network services						$200,000		$350,000		$500,000		$500,000		$500,000		$500,000		$500,000		$500,000	$3,550,000
o Computer hardware						$250,000		$200,000		$150,000		$150,000		$150,000		$150,000		$150,000		$150,000	$1,350,000
o Computer software						$100,000		$100,000		$50,000		$50,000		$50,000		$50,000		$50,000		$50,000	$500,000
Subtotal, NVO operations						$730,000		$1,010,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000		$1,240,000	$9,180,000
Surveys and Data Access Grants Program		$300,000		$500,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000		$750,000	$6,800,000
Research Grants				$500,000		$800,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000		$1,500,000	$11,800,000
Fellowship Programs				$300,000		$500,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000		$1,000,000	$7,800,000
Education and Public Outreach			1	$150,000	1	$150,000	2	$300,000	3	$450,000	4	$600,000	4	$600,000	4	$600,000	4	$600,000	4	$600,000	$4,050,000
Management					2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	2	$440,000	$3,520,000
Administrative					1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	1	$75,000	$600,000
Totals	19	$3,420,000	26	$5,420,000	28	$6,845,000	34	$8,835,000	37	$9,275,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	31	$8,365,000	$75,620,000
Totals, Inflation Adjusted		$3,420,000		$5,610,000		$7,333,000		$9,796,000		$10,643,000		$9,935,000		$10,283,000		$10,643,000		$11,015,000		$11,401,000	$90,079,000
Integral Cost		$3,420,000		$9,030,000		$16,363,000		$26,159,000		$36,802,000		$46,737,000		$57,020,000		$67,663,000		$78,678,000		$90,079,000	
Year		2002		2003		2004		2005		2006		2007		2008		2009		2010		2011	

Rates and assumptions:
Admin support labor rate, fully burdened $75,000

Operations staff labor rate, fully burdened $90,000
Technical labor rate, fully burdened $150,000
Management labor rate, fully burdened $220,000
Inflation rate   1.035

70

Appendix F: The Team Membership

Chairman:

George Djorgovski (Caltech), george@astro.caltech.edu

Regular Members:

Charles Alcock (U. Penn.), alcock@hep.upenn.edu
Piero Benvenuti (ESO), pbenvenu@eso.org
Roger Brissenden (CfA), rjb@head-cfa.harvard.edu
Derek Buzasi (USAF Academy), Derek.Buzasi@usafa.af.mil
Dave DeYoung (NOAO), deyoung@noao.edu
Isabel Hawkins (UC Berkeley), isabelh@ssl.berkeley.edu
George Helou (Caltech/JPL/IPAC), helou@ipac.caltech.edu
Frank Hill (NSO), hill@noao.edu
Stephen Kent (Fermilab), skent@fnal.gov
Paul Messina (Caltech), messina@cacr.caltech.edu
Andrew Moore (CMU), awm@cs.cmu.edu
Jim Schombert (U. Oregon), js@abyss.uoregon.edu
Alex Szalay (JHU), szalay@pha.jhu.edu
Meg Urry (Yale), meg.urry@yale.edu
Nicholas White (NASA GSFC), nwhite@lheapop.gsfc.nasa.gov

Consultative Members:

Robert Brunner (Caltech), rb@astro.caltech.edu
Pepi Fabbiano (CfA, ADCCC), pepi@head-cfa.harvard.edu
Eric Feigelson (Penn State U.), edf@astro.psu.edu
Francoise Genova (CDS), genova@newb6.u-strasbg.fr
Jim Gray (Microsoft Research), gray@microsoft.com
Jon Hakkila (College of Charleston), hakkilaj@cofc.edu
Bob Hanisch (STScI), hanisch@stsci.edu
Sally Heap (NASA GSFC), Sara.R.Heap.1@gsfc.nasa.gov
Roberta Humphreys (U. Minn.), roberta@anubis.spa.umn.edu
Barry Madore (IPAC/NED), barry@ipac.caltech.edu
Roger Malina (UC Berkeley), rmalina@prontomail.com
Janet Mattei (AAVSO), jmattei@aavso.org
Tom McGlynn (NASA GSFC, USRA), tam@silk.gsfc.nasa.gov
Robert Nichol (CMU), nichol@cmu.edu
Ethan Schreier (STScI), ejs@stsci.edu
Mark Sykes (U. Arizona), sykes@as.arizona.edu

Ex Officio:

Joe Bredekamp (NASA HQ), joe.bredekamp@hq.nasa.gov
Wayne Van Citters (NSF AST), gvancitt@nsf.gov
Eileen Friel (NSF AST), efriel@nsf.gov

Appendix G: 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/

Appendix H: Glossary of Acronyms

2MASS	Two Micron All Sky Survey
AASC	Astronomy and Astrophysics Survey Committee
ADC	Astronomical Data Center
ADEC	Astronomical Data Center Executive Council
ADP	Astrophysics Data Program
ADS	Astrophysics Data System
AGN	Active Galactic Nuclei
AISRP	Applied Information Systems Research Program
AO	Announcement of Opportunity
ASCA	Advanced Satellite for Cosmology and Astrophysics
AVO	Astrophysical Virtual Observatory
CCD	Charge Coupled Device
CfA	Center for Astrophysics
CFHT	Canada France Hawaii Telescope
CMBR	Cosmic Microwave Background Radiation
COMRAA	Committee on Organization and Management of Research in Astronomy and Astrophysics
CXO	Chandra X-ray Observatory
CY	Calendar Year
DARPA	Defense Advanced Research Projects Agency (DoD)
DOE	Department of Energy
DPOSS	Digital Palomar Observatory Sky Survey
DTF	Distributed Terascale Facility (NSF)
EMSS	Einstein Medium Sensitivity Survey

EPO	Education and Public Outreach
EGSO	European Grid of Solar Observations
FITS	Flexible Image Transport System
FY	Fiscal Year
GO	Guest or General Observer
GONG	Global Oscillation Network Group
GRB	Gamma Ray Burst
GSRP	Graduate Student Research Program
HST	Hubble Space Telescope
IRAS	InfraRed Astronomical Satellite
IT	Information Technology
ITR	Information Technology Research
JPL	Jet Propulsion Laboratory
K-12	Kindergarten through 12th Grade
LSST	Large-aperture Synoptic Space Telescope
MACS	Massive Cluster Survey
NAS	National Academy of Sciences
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputing Applications
NED	NASA Extragalactic Database
NGST	Next Generation Space Telescope
NSF	National Science Foundation
NVO	National (U.S.) Virtual Observatory
OADS	Original Astrophysics Data System
OSS	Office of Space Science
QUEST2	Quasar Equatorial Survey Telescope
RDBMS	Relational DataBase Management Systems
RDCS	ROSAT Deep Cluster Survey
REU	Research Experiences for Undergraduates
ROSAT	Röntgen Satellite
SAWG	Science Archives Working Group
SDAC	Solar Data Analysis Center
SDSC	San Diego Supercomputing Center
SDSS	Sloan Digital Sky Survey
SDT	Science Definition Team
SETI@home	Search for ExtraTerrestrial Intelligence @home
SHARC	Serendipitous High-redshift Archival ROSAT Cluster Survey
SIMBAD	Set of Identifications, Measurements and Bibliography for Astronomical Data
SMM	Solar Maximum Mission
SOHO	Solar and Heliospheric Observatory
SOLIS	Synoptic Long Term Investigation of the Sun
STScI	Space Telescope Science Institute
TRACE	Transition Region and Coronal Explorer
VIRMOS	VIsible imaging Multi-Object Spectrograph
VLA	Very Large Array
VLT	Very Large Telescope
VO	Virtual Observatory
VSO	Virtual Solar Observatory
WARPS	Wide Angle ROSAT Pointed Survey
WWW	World Wide Web
XML	eXtended Markup Language
XMM	X-ray Multi-Mirror Satellite