

ITR/IM:
Building the Framework for the National Virtual Observatory

Proposal to the National Science Foundation
Information Technology Research Solicitation
23 April 2001

Principal Investigators

Paul Messina
California Institute of Technology

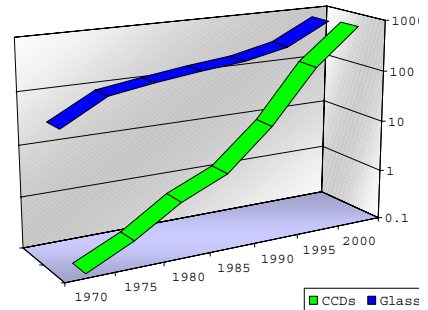
Alex Szalay
Johns Hopkins University

Senior Personnel

Charles Alcock	University of Pennsylvania
Kirk Borne	Astronomical Data Center/Raytheon
Tim Cornwell	National Radio Astronomy Observatory
David DeYoung	National Optical Astronomy Observatories
Giussepina Fabbiano	Smithsonian Astrophysical Observatory
Alyssa Goodman	Harvard University
Jim Gray	Microsoft Research
Robert Hanisch	Space Telescope Science Institute
George Helou	NASA Infrared Processing and Analysis Center
Stephen Kent	Fermilab
Carl Kesselman	University of Southern California
Miron Livny	University of Wisconsin, Madison
Carol Lonsdale	NASA Infrared Processing and Analysis Center
Tom McGlynn	GSFC/HEASARC/USRA
Andrew Moore	Carnegie-Mellon University
Reagan Moore	University of California, San Diego
Jeff Pier	United States Naval Observatory, Flagstaff Station
Ray Plante	University of Illinois, Urbana-Champaign
Thomas Prince	California Institute of Technology
Ethan Schreier	Johns Hopkins University/STScI
Nicholas White	NASA Goddard Space Flight Center
Roy Williams	California Institute of Technology

1 Scope and Vision of the NVO

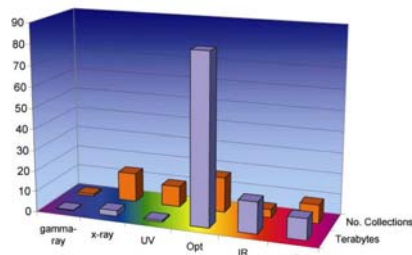
Astronomy faces a data avalanche. Breakthroughs in telescope, detector, and computer technology allow astronomical surveys to produce terabytes of images and catalogs. These datasets will cover the sky in different wavebands, from gamma- and X-rays, optical, infrared, through to radio. In a few years it will be easier to “dial-up” a part of the sky than wait many months to access a telescope. With the advent of inexpensive storage technologies and the availability of high-speed networks, the concept of multi-terabyte on-line databases interoperating seamlessly is no longer outlandish [NVO][EUUS]. More and more catalogs will be interlinked, query engines will become more and more sophisticated, and the research results from on-line data will be just as rich as that from “real” telescopes. Moore’s law is driving astronomy even further: the planned Large Synoptic Survey Telescope will produce over 10 petabytes per year by 2008! These technological developments will fundamentally change the way astronomy is done. These changes will have dramatic effects on the sociology of astronomy itself.



The total area of astronomical telescopes in m², and CCDs measured in Gigapixels, over the last 25 years. The number of pixels and the data double every year.

On-line astronomy demands new IT approaches that will yield tools and methodologies for data access, analysis, and discovery that are scalable to this regime. New needs lead to opportunities in IT research for data mining, for sophisticated pattern recognition, for large-scale statistical cross-correlations, and for the discovery of rare objects and sudden temporal variations. With a billion objects, statistical algorithms requiring N^3 steps would take billions of processor-years—even $MlogN$ algorithms will take a long time, creating challenges in their own right! Moreover, there is a growing awareness, both in the US 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 demand efficient and effective synthesis of these capabilities—both for astronomy and for the broader scientific community.

Recognizing these trends and opportunities, the National Academy of Sciences Astronomy and Astrophysics Survey Committee, in its decadal survey [NAS99] recommends, as a first priority, the establishment of a **National Virtual Observatory**. The NVO would be a “Rosetta Stone”: linking the archival data sets of space- and ground-based observatories, the catalogs of multi-wavelength surveys, and the computational resources necessary to support comparison and cross-correlation among these resources. The NVO will benefit the entire astronomical community. It will democratize astronomical research: the same data and tools will be available to students and researchers, irrespective of geographical location or institutional affiliation. The NVO will also have far-reaching education potential. Astronomy occupies a very special place in the public eye: new discoveries fascinate both the large number of amateur astronomers and the general public alike. The NVO will be an enormous asset for teaching astronomy, information technology, and the method of scientific discovery. Outreach and education will be key elements: the NVO will deliver rich content via the Internet to a wide range of educational projects from K-12 through college and to the public.



The chart shows the sizes and the numbers of the collections participating in this proposal by their wavelength (see Appendix A).

This proposal describes the NVO’s scientific opportunities and Information Technology challenges. It presents an implementation strategy and management plan to create an initial federation, and a foundation for further tools and applications. The NVO will challenge the astronomical community with new opportunities for scientific discovery and it will challenge the information technology community with a visionary but achievable goal of distributed access and analysis of voluminous data collections. The scope of the effort is international.

By its very nature the NVO brings together groups with different talents. Our proposal recognizes that the ability to fully exploit the federation of all major astronomy archives as an integrated resource requires

cooperation. Our team represents an unusually wide range of astronomical and IT expertise. It includes data providers from space missions, ground-based telescopes, and special surveys, from leading astronomical institutions throughout the country. Our team members have committed to federate more than 100 Terabytes of astronomical data, consisting of over 50 collections in the NVO. The team has strong IT expertise in advanced networking applications, management of very large databases, Grid computing [FK99], data mining, and advanced statistics.

2 The NVO Science and IT Paradigm

2.1 Enabling New Science—The New Astronomy

In order to be an *engine of discovery for astronomy* and enable qualitatively new advances, the NVO must be driven by science goals. We think of the NVO as a genuine observatory that astronomers will use from their desks. It must supply digital archives, metadata management tools, data discovery, access services, programming interfaces, and computational services. Astronomers may develop their own custom programs to answer specific questions, sifting through the vast digital sky to identify rare objects, compare data with numerical models, and make discoveries through advanced visualizations and special statistical analyses. In addition, students, teachers, under-represented groups, and the general public have equal access to this cutting-edge resource from tailored portals.

The importance of large, uniform samples of multi-wavelength and temporal information about astronomical objects has been amply demonstrated. This information will come from different catalogs residing in different parts of the NVO federation. While much of this proposal concerns IT, we first select a few cutting edge astronomy topics and discuss the NVO's scientific impact on them.

Comparing the Local and the Distant Universe: Combining IR and optical observations has opened a new window on the distant universe. Having a broad range of colors for distant galaxies enables us to estimate not only photometric redshifts, but also to spectral type, and study detailed star formation history. The NVO will let us create rest-frame selected samples from combined UV, optical, and IR datasets. Comparing local and distant samples, we can study the evolution of physical properties, such as the infrared-radio correlation in star-forming galaxies. Statistical analyses of samples defined via multi-wavelength queries in the NVO will measure spatial clustering patterns as a function of redshift, revealing density fluctuations in the early universe and constraining values of the fundamental cosmological parameters. Massive cluster surveys will enable us to trace the complex evolution of large-scale structure, from its origins in the cosmic microwave background to the amazing diversity we see today; MAP and PLANCK promise views of the infant universe at high resolution. X-ray missions (Chandra, XMM, ROSAT), and deep ground based observations show structure to $z \sim 4$, and we will later probe to $z \sim 30$ with NGST. SIRTf and other IR surveys will fill in the gap at intermediate redshifts. Distortions in optical galaxies around these clusters provide direct measures of dark matter and its distribution. Using the NVO, automated detection of odd-shaped objects (arcs) in petabytes of imaging data (both optical and radio) will enable us to find many strong gravitational lenses.

Digital Milky Way: We know surprisingly little about the origin and evolution of our own galaxy. Federating all existing information on the Milky Way will enable systematic mining of multi-wavelength catalogs and surveys, both existing and yet to be created. The Milky Way galaxy is a complex entity consisting of multiple stellar components (bulge, disk, halo), each with its own mass function, metallicity, age, and kinematics. The interstellar medium is equally complex. Current surveys covering 100 square degrees already reveal halo kinematic substructures in the form of star streams and accretion remnants. Current surveys capture all halo giants out to 100 kpc and all but the faintest dwarfs to 15 kpc: deeper surveys will allow for the first time a definitive study on the origins of the Galactic halo and thick disk. An important challenge, for example, is to understand the role of fast encounters with other galaxies (typically smaller than the Milky Way) in the evolution of our Galaxy, which can create new stellar components, impact the evolution of existing components, and can directly induce star formation. The planned Large Synoptic Survey Telescope (LSST) will provide high precision proper motions from co-added images that will help address these crucial problems. These measurements will also provide an ideal calibration of large-scale n-body and hydrodynamic simulations of galaxy encounters. The data produced by the simulations will themselves reside in databases, and can be “observed” through the same tools as the rest of the NVO.

Rare and Exotic Objects: Large surveys detect significant numbers of outliers in the statistical distributions of derived parameters. These anomalies are often not immediately obvious from a single observation; and are not followed up scientifically, since the discoverers may not be able to devise a compelling model for the phenomena using the limited data at their disposal. The NVO will enable astronomers to find objects that can only be identified

by being statistically unusual when multiple-wavelength catalogs are compared. Early multicolor surveys (SDSS, 2MASS, DPOSS2) have led to the discovery of distant quasars and brown dwarfs. The search for new classes of objects in far larger volumes of parameter space will remain untouched until the NVO is created. The search for rare objects in the temporal domain could yield some of the most exciting new results from the NVO: rapid identification of transient objects by comparing new observations with prior epochs in real-time may reveal distant supernovae and gravitational microlensing. The side-product will be an unprecedented catalog of variable stars. Other projects include the search for possible dark matter constituents in the Halo of our own galaxy (e.g., cool white dwarfs), near Earth asteroids, unusual variability in “known” source classes, as well as new types of transient sources. Does the sun have a binary companion? Surprisingly, this is possible and the answer is not yet known: it could be an under-luminous dwarf in a very elongated orbit. A systematic search through a combination of large sky surveys in the optical and near-IR taken over a large time baseline could answer this intriguing question.

Census of Active Galactic Nuclei (AGN): There has been a major shift in our understanding of the role of supermassive black holes in galactic formation and evolution. These black holes have masses in the range of 10^6 - 10^{10} solar masses and we now believe that most galaxies harbor such black holes at their centers. AGN are “beacons” which signal the presence of a black hole in a galaxy, shining by converting the energy from accreting matter into radiation. Much of the radiative output of the early universe may have been emitted by these accreting supermassive black holes. However, despite decades of effort, a census of AGN and their place in the galactic evolutionary scheme still eludes us. The chief problem is dust obscuration: hidden AGN may outnumber their unobscured counterparts by an order of magnitude. Glimpses of this population are now being seen in x-rays. The NVO will include the deepest X-ray data ever obtained (Chandra); the most extensive redshift catalogs (SDSS); the highest resolution optical observations (Hubble); the largest infrared archives (2MASS, then SIRTf); the high-resolution spectra from the VLT and Gemini; radio data from the Very Large Array; and a vast panoply of planned future surveys at various wavelengths. It will enable a panchromatic census of AGNs, allow us to probe the connection between galaxies and supermassive black holes, and reveal the cosmic history of energy production from both nucleosynthesis (star formation activity) and accretion (AGNs).

Search for Extra-Solar Planets: The search for extra-Solar planets is a major goal of twenty-first century astronomy; it carries with it the scientist’s hunger for new understanding, the philosopher’s inquiry into the meaning of life, and the public’s desire to know the answer to a simple question: “are we alone?” The spectacular recent progress with the radial velocity technique has already overturned the standard models for planet formation, but also revealed its frustrating limitations. Other techniques are being pursued: the planet-transit technique readily scales to surveys around very large numbers of stars, and would benefit enormously from the federation of data from multiple future surveys. In particular, the data taken by the Large Synoptic Survey Telescope, coupled with infrared surveys (2MASS and its successors) and astrometric surveys (FAME), will enable a survey for planetary transits around billions of stars in the Milky Way and in several nearby galaxies. This survey would go beyond the parochial vision (“do we have near neighbors?”) to the larger questions, such as “how many planets are there in the universe?”

Theoretical Astrophysics: The NVO will make possible, for the first time, truly significant interactions between large datasets and the equally large-scale theoretical simulations of astrophysical systems that are just now becoming available. In a few years, use of massively parallel terascale computing systems will allow: a) the calculation of the orbits and evolution of every star in a globular cluster, including stellar collisions, the formation of tight binary systems, core evolution, and the effect of all these on the stellar evolutionary tracks; b) the details of galaxy encounters and mergers, including both the fate of the stars and the interstellar medium; c) the evolution of the large scale structure of the Universe, including the formation of galaxies, clusters of galaxies, and clusters of clusters. All of these and similar theoretical calculations will produce datasets comparable in size with those of the large scale observational surveys, and it will be possible to mine these datasets just like observational data. Definitive comparisons will be made between complex theoretical calculations and observational datasets large enough to be statistically significant in all parameters. These studies, which will be carried out within the framework of the NVO, will lead to solutions of some of the most outstanding and significant astrophysical problems of our time.

2.2 The NVO as a Semantic Web

The primary focus of the NVO is data federation, fusion, and exploration. In order to achieve these goals the NVO will break new ground as a large-scale prototype of a *semantic web* [BL01]. The NVO will federate a large number of heterogeneous data sets distributed around the world. Some of these data sets are small, some are tens of terabytes in size, some are under database management, and others are not. The data include catalogs of objects

with attributes, image data of varying resolution and wavelength, spectral data, temporal data, and ancillary reference literature. An ambitious goal of the NVO is to federate the data and information of an entire scientific discipline. Computational grid technologies will provide access to distributed computing resources that enable the creation of the terascale analysis pipelines that will be required for some investigations.

The NVO team consists of a close collaboration between computer scientists and application scientists, and the experience gained from development of the NVO can be expected to significantly impact Information Technology research in the future. Astronomy provides an ideal environment for developing a large scale data grid prototype: the community of astronomers is moderate in size (a few thousand) but not too large; the data sets are heterogeneous, but not exceedingly so, providing interesting, tractable challenges for metadata standards and protocols; the data repositories are widely distributed, but typically already electronically accessible; security is not usually emphasized in astronomy; and finally, astronomy is of widespread interest to the public, providing an interesting proving ground for a knowledge network that engages many levels of society.

2.3 The NVO as Facility for Data Publication

A long-standing problem in astronomy and other sciences has been how to publish data in addition to a scientific paper. The advent of the Web has shown us how easy it is to “publish” a web page, and we intend to make data publishing just as easy for astronomers, yet in a semantically-rich fashion that allows readers to find it, read it, assess its provenance, compute with it: in other words, to make full use of it.

The NVO must recognize the three different roles of *author*, *publisher/curator*, and *reader*. Data authors will work with publishers to “publish” their data, i.e., to provide full digital access to information which has traditionally only been available in graphs or tables in printed papers. When this process is complete, the data moves to the archive along with the metadata and documentation. In addition, a “standard” form of the data will be generated: we need to translate measurements into standard units wherever possible, translate data into the standard representations supported by the archive (data models), and perhaps define a few new measurements and representations. Publishing astronomical data must become far less difficult than it is today. It is the task of the NVO to develop the tools, templates and standards that make it easy to document and publish data, and to make it cost-effective for the archivists to manage and curate published data. In the NVO era an astronomer wishing to publish data will be able to characterize it through an NVO “publication portal” which captures metadata describing the content in NVO compliant terms, identifies the access mechanisms, and registers the resource with the NVO. In this process the NVO publishing standards do not obscure the raw data; users will always be able to “drill down” to data in the original format. The standardization is used only for locating and federating disparate data sources.

3 The Main Challenges

Meeting the NVO’s unique IT challenges will both enable new science *and* advance our IT technologies into a petascale data grid, soon to be a frontier for US business as well as science. Knowledge extraction from billion-object catalogs requires new indexing and summarization techniques. Petascale pixel image analysis from multiple distributed archives will require integration of digital library and grid middleware. As new classifications are discovered, understood, and archived, the NVO catalog will have to evolve. A *data management system* will provide a uniform data access layer for data pipelining, archiving and retrieval of terabytes of distributed astronomical images. An *information management system* will support inserting, querying, and evolving billions of objects each with thousands of attributes. A *knowledge support system* will provide software tools for correlation, visualization, and statistical comparisons of both cataloged data and original image pixel data.

The NVO architecture will be based on middleware that integrates federated, distributed, autonomous archives. It will connect users to analysis services and data services. The analysis-oriented service will support massive data analysis of catalog and pixel image data. The key functional requirements for the middleware are:

Handling distributed collections and resources. The NVO will federate existing astronomical data into a cooperating system. As such, interoperability and autonomy are key attributes of any design. Each participant must expose a uniform data access layer, but the internals of each member of the federation will quite likely remain unique. Collection integration will require mediation across the diverse semantic conventions used to describe all wavelengths and both ground- and space-based sensors.

Providing a uniform astronomy information infrastructure. The NVO will maintain international links to similar efforts in other countries, and in other disciplines. We will work to promote international standards for data and information access. This will lead to direct enhancement of international collaborations in astronomy.

Enabling tera/petascale data analysis. The NVO will provide integrated access to terascale computing and data facilities in a way that is useful to astronomers. The emphasis is on seamless scalability from desktop to supercomputer, so that what is learned at one stage need not be unlearned later.

Capitalizing on advancing technology. We will incorporate commercially available technology where possible, while developing the needed large-scale statistical analysis and image analysis capabilities that are not currently available. We will focus on interoperability and open standards, and plan to adapt as new technologies emerge.

Data federation and fusion is a prime focus of NVO. 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 data, there are additional requirements. Caching and replication services can save the results of complex joins of multiple databases for later reuse. An efficient proximity join of a billion sources requires that the data are clustered so that nearby objects in the sky are nearby in the stream: we plan to use the HTM indexing developed at Johns Hopkins to achieve this [KST01].

Users have the most control and interactivity with their desktop workstations, so small datasets will probably be brought to the desktop for visualization and for experimentation. However, the initial stages of the pipeline may involve huge data volumes distributed over a wide area algorithms must be moved to the data. Thus agent-code portability is as important as the portability of the data format. While many operations can be controlled by menus and numerical parameters sophisticated users will want to write compiled code that can execute near the data.

The NVO framework will work toward widely accepted astronomical *metadata standards and protocols*. These must be extensible into the far future. XML will be our fabric for structured information, including interoperation of between Astronomical XML, Astronomical Markup Language [GUI98], Astrores [ACC99], Extensible Scientific Interchange Language [XSIL], and other astronomical data representations. FITS is a standard for structured data that predates XML, and a first milestone of this proposal is a software toolbox for FITS/XML interoperation.

The NVO will help extend the “profiles” defined by NASA’s Space Science Data System, and its prototype implementation in the Astrobrowse system. In addition to the interfaces among web services, we must extend the metadata semantics so that programs can be written against a “metadata API”. Other, non-astronomical, objects must also be described for a successful NVO, and we hope to borrow from other projects and from the commercial world for adequate semantic and syntactic descriptions. Examples are Document, Published article, Preprint, Person, Link, Parameter, Array, Image, Message, Exception report, Service description.

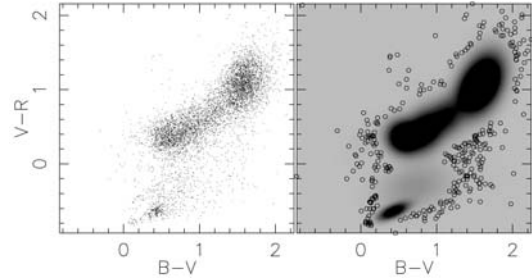
Scalability is a major challenge: The NVO must handle billions of objects in high-dimensional petascale astronomical catalogs. It must also scale to an international federation of hundreds of institutions: some huge and some tiny. The computational grid will help by providing massive online storage, parallel computing, resource management, and high-speed network access, but the NVO must solve the problems of data organization, data access, data analysis, and data visualization. We will attack these problems by a combination of cunning and brute force: where possible the NVO will build sophisticated indices, pre-compute popular aggregates, and cluster the data for efficient access. But, in the end, the curse of dimensionality, or the ad hoc nature of some queries will force bulk data scans. In these cases, the NVO will use brute force: both providing very high speed sequential access to the data, and providing compact replicas of the data (bitmaps or tag objects) that minimize the amount of data that must be scanned to answer a query. Of course, all these operations will be done in parallel using both parallel processing and parallel IO. This parallelism should come from the data management tools, but if required, the NVO will implement these mechanisms.

In the end, the system will be judged by how quickly users can pose questions, and understand the answers. This means that the system will be responsible for translating high-level non-procedural queries into efficient execution plans, executing the plans so as to minimize data movement, and then deliver the results to the visualization tool as quickly as possible, perhaps allowing the user to steer the computation as it progresses.

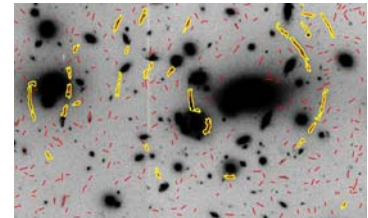
The diverse user communities will access the NVO through networks of varying speeds; the tools and human interfaces must be usable for both low and high-speed connections. Some analyses will require very-large-memory machines and computing speeds, while others will fit well on Beowulf class systems. The computational grid and the NVO federation will include and support both computation styles.

NVO will serve a large user community. The number of astronomers worldwide is only a few thousand; however, the education and public access functions will have user communities numbered in the millions. While each such user will impose a modest load, the aggregate will be substantial. The NVO framework must be designed to handle large numbers of small users as well as a few very large users.

Data Understanding: Things are changing in the way that data analysis takes place in astronomy: intensive use of algorithms on data is replacing the personal astronomer/data relationship of the previous generation. The NVO will accelerate this process, so that thousands of astronomers can benefit from the data without drowning in it. There will be many astronomers wishing to analyze the data in many different ways. We will provide tools and algorithms that support statistical and data mining queries needed by astrophysicists, and interface these to the framework. These include spatial access methods such as kd-trees, R-trees, metric trees and newer generations thereof, and also condensed representations such as sparse datacubes and binned grids. These structures support queries regarding n-point correlations and non-parametric density estimates. We will provide example implementations that will serve as well-documented tutorials on interfacing one's own analysis algorithm with the NVO and will also provide directly useful tools. Examples of such analyses involve the identification of rare objects, or automated shape finders searching for atypical objects, like the gravitationally lensed arcs.



Automated identification of outliers in the multicolor distribution of SDSS stellar sources, using the EM algorithm. This density estimation technique improves the efficiency of quasar detection from 65% to 90%, using 5 color SDSS data [CON01].

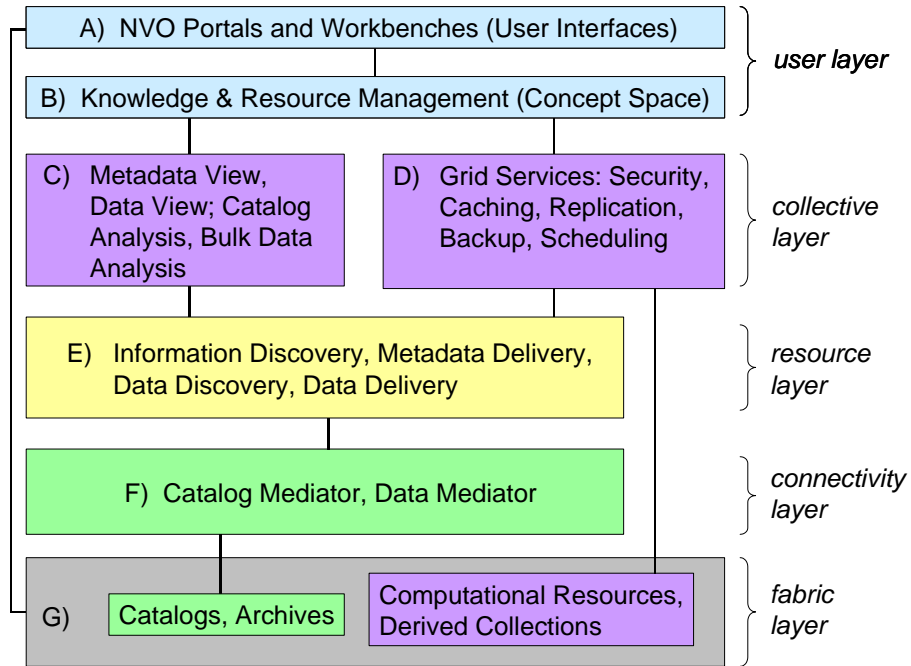


A prototype of an automated discovery tool for gravitationally lensed arcs [SZ01]. Shown is a Hubble image of the cluster A2218. Such automated procedures could be run on all Hubble images and discover large numbers of new lenses.

4 NVO Architecture

The NVO will enable new astronomical research by developing mechanisms to federate collections of data, publish new collections, access images, compare catalogs, and perform image and statistical analyses that require processing large fractions of the collections. Our infrastructure builds upon capabilities from grid environments for high performance distributed analysis, data grid environments for data, information, and knowledge management, and digital libraries for data discovery, analysis, and presentation. We meet the challenge of integrating technology from all three communities, while providing access to existing astronomical archives with unique data and metadata access mechanisms. The NVO is not an effort to integrate all astronomical services via top-down control: its goal is to provide bottom-up frameworks and toolkits to make these services integrable on whatever scale is appropriate for specific user needs. At the simplest level, the influence of NVO may be as subtle as adjusting result-set formatting toward existing data standards. At the high end, it will define distributed processing command/control methodologies. The NVO framework will enable and foster integration but not mandate it.

The NVO architecture is based upon the standard Grid differentiation between the *fabric layer* (compute and storage platforms), the *resource layer* (standard mechanisms for data and metadata discovery and delivery), and the *collective layer* (federated data and distributed computing). The architecture is augmented with two layers specifically designed to address NVO interoperability requirements, a *connectivity layer* that translates between individual archive access protocols and the NVO grid access protocol, and a *knowledge repository* that manages the NVO concept space that is used to define relationships between attributes used within different data collections. The approach relies upon the development of interoperability mechanisms to minimize the effort required for collections to join the NVO federation. This approach also allows for multiple entry points—*portals and workbenches*—designed for specific user groups or functions (astronomers with different discipline interests, educators, students, the general public, etc.).



The correspondence of the NVO architecture layers to the Grid infrastructure layers is shown on the right side of the diagram. Each component is designed to support access to the existing survey digital libraries and to the expanded capabilities required by the NVO to support analyses that require processing of a large fraction of the catalog holdings or images from multiple surveys.

4.1 The User Layer

The user layer provides the highest level interfaces between the user and the NVO: portals and workbenches. The latter depend upon *knowledge management* services to aid data understanding, including global data discovery services integrating metadata from many data collections, and upon *resource management* services to optimize and schedule large scale distributed computations, including managing replicated datasets or entire data collections.

Portals and Workbenches

Astronomy already provides many web-based data portals, as well as sophisticated data processing workbenches, and we do not wish to duplicate these within NVO. We propose instead to foster interoperability, flexibility, and extension of these existing portals, while adding a new portal to enable Grid-based supercomputing. Existing web portals support data download and computing workbenches that run on a single CPU. The NVO vision, however, includes massive computation on massive datasets, organized as a distributed data pipeline. A user might want to generate spectra for objects seen in a set of Chandra images, using an object detection tool provided by the Chandra archive, and a spectral analysis tool installed at HEASARC. To address such a problem the user will need a toolkit to create and manage the necessary tasks: select, initiate, monitor, steer, kill, define input/output, etc.

We will create a portal to allow submission of such jobs. Each node of the pipeline will wait for its I/O resources to become available before beginning computation, in a dataflow paradigm. The portal will be visible in equivalent versions as a command-line script, or as a boxes-and-pipes GUI. For performance, we can allow each node to be an SPMD parallel computer. Similarly, each connector can be a parallel stream, where the order of receipt of messages may be different from the order of sending. The portal will be integrated with the Grid infrastructure currently being developed by the GriPhyN collaboration [GRIP], including scheduling, caching, and debugging.

We will build a powerful system for visualization by leveraging hardware funded by other NSF grants. The computational component will process complex selection queries from a large catalog and run sophisticated clustering and other analysis software, transforming the data into drawing instructions for an immersive 3D visualization system. The investigator will manipulate the transformations in real time, creating a rich investigation

environment in this purely abstract space. All of the software developed for this project will be portable to desktop machines, so that investigators can learn effectively and trust the system. Geometric representations will, to the greatest possible extent, be of graduated resolution; as the picture gains focus gradually; an investigator will be able to make decisions before the rendering is complete.

Knowledge Management Services

NVO will need services to share knowledge across the entire system. We will define a semantic concept space matching collection attributes to NVO concepts, and a service interoperation space, composing services across collections, including data reformatting and unit conversion.

The most fundamental NVO knowledge service is global *data discovery*. The resource layer metadata-handling interface will be used to download and integrate metadata from many data collections to support user-oriented (science-attribute-driven) queries to locate data of interest. Service-provider-defined *profiles*, which characterize the nature of a data or service resource and the metadata attributes of its holdings, will be used to search for data matching user-defined selection criteria. Within astronomy, research efforts such as Astrobrowse, ISAIA [HAN00], and CDS GLU have prototyped the use of profile-based schemes for data discovery. We will use XML formats for this modeled on the Digital Earth program [DE] and the Web Service Description Language [WSDL].

In a dynamic, widely distributed system like the NVO, data discovery will be hierarchical in nature; no single data discovery service will index all data. Rather, regional or discipline-specific centers will index all known data of a certain type, and the largest NVO nodes will index data from multiple smaller centers as well as from individual data collections. The system will support “living archives” subject to frequent updates as new data is added or as existing data is modified. Other knowledge services, many of which already exist within astronomy, will need to be integrated with the NVO. These include services for name resolution (NED, SIMBAD), mapping of astronomical object names to positions on the sky, and cross-referencing data objects with published papers.

Resource Management Services

The NVO framework must support computation that is scaleable from user workstation-based analysis of downloaded datasets to bulk catalog or image analysis of massive replicated data collections on supercomputers. *Replication* of datasets or entire data collections will be employed to cache data to optimize computation while providing data backup. Replica management services, integrated with data discovery, will be used to perform replica selection to stage data and optimize distributed computations. Request planning and scheduling services which combine resource level information on computational resources and network bandwidths with replica selection performance metrics will be used to plan and execute large distributed computations. We will extend the Storage Resource Broker (SRB) to provide collection-based bulk data replication services as well as transparent access to data stored in a network data cache (*e.g.*, a GridFTP server, DPSS, or a Unix file system) or in tertiary storage such as HPSS. Technologies such as CONDOR will be used to provide supercomputer-class computation for data-parallel problems using large collections of commodity workstations.

4.2 The Collective Layer

The collective layer of the NVO architecture builds upon the capabilities provided by multiple distributed resources to provide *computation* for large-scale statistical data analysis and queries, large scale-catalog cross-correlations, and bulk processing of very large image databases; and *analysis* web services for multi-wavelength astronomical data analysis and data fusion.

Metadata and Data View

We will provide an interface to the metadata profile published by a data-provider—a human-readable version of the metadata profile, with short and long definitions of each term, attribute, or keyword. We will also build an API to the metadata profile library and implement bindings in popular languages. An “XML toolkit” will be customized for NVO so that astronomers’ programs can effectively utilize and generate the XML-based messages that will form the backbone of NVO communication. Drawing components will use standards such as VML (Vector Markup Language), or image-processing services such as a contour plot from an image. Catalog components and image components will be able to display, select, and combine these objects, as well as write them in different formats. Directory components will convert a query on a directory into a set of object references. Navigation components will allow browsing based on object type, sky position, and bibliographic references.

Catalog and Bulk Data Analysis

With more structured data services will come the ability for third parties to construct simple meta-services. These services will rely on the more basic building blocks (such as those listed above) to construct focused processing scenarios that then become services in their own right. These integrating services may coordinate other services or may directly reference specific datasets. Examples include distributed relational joins (catalog to position-based cross-referencing information to catalog, all three potentially from different sources), and image mining (source extraction, feature measurements, aperture photometry) based on source lists (catalog subsets).

Grid Services

User demands for service access and data processing will always outstrip available resources. NVO will therefore include capabilities for resource allocation, process scheduling, and accommodation of delays between request and response. While NVO sites participate with the intention of increased sharing of astronomy assets, such sharing may not be unconditional. Furthermore, the NVO must be able to control access to resources such as storage and computation. To do so, the NVO must know the identity of the data requestor. Control must be uniform across all participants in the federated collection of resources participating in the NVO.

We will develop a single sign-on environment where a user authenticates once to the NVO and can then access any resource that they have rights to use. Once user identity is established, community authorization servers can be used to establish and enforce usage policy. Both authentication and access control technologies will be based on the security infrastructure of the Globus toolkit. These security solutions are widely deployed, and their use will ensure NVO interoperability with other Grid resources, notably those being deployed by the NSF Partnerships for Advanced Computing Infrastructure (PACIs), the GriPhyN project, and NASA's Information Power Grid.

4.3 The Resource Layer

The resource layer components of the NVO architecture define the basic services, protocols and mechanisms from which all higher-level NVO functionality is built. This layer forms the boundary between local management issues and more global NVO specific issues. The resource layer provides mechanisms for: *discovering* the location and characteristics of NVO resources: computational elements, storage systems (e.g., information services), and data items of interest (e.g., metadata); *providing access* to data elements, either one at a time, or in bulk; and *initiating, monitoring, and managing* data-analysis computations, either close to the data or at a remote location. The resource layer does not encompass higher-level services (e.g., for data discovery or multi-wavelength data fusion), which may combine data from multiple archives. A single site may however provide both data access and resource management as well as higher-level services. It is vital that the NVO builds upon and coordinates with related efforts in the Grid, Data Grid, GriPhyN / PVDG, and XML / Digital Library communities. We will adapt or extend existing or developing Grid technology, such as the Globus toolkit and the Storage Resource Broker, for most of the framework-oriented parts of NVO services (Grid Collective) layer, e.g., for dataset replication, replica management, information discovery, resource discovery and request management.

Resource and Information Discovery

The NVO is a loosely coupled federation of data collections, users, and resources, and participation will change over time. There need to be mechanisms by which the resources available in NVO at any time can be discovered and queried. We propose to use directory services based on the Globus information services, and also on the emerging UDDI protocol [UDDI].

While the service discovery mechanisms discussed above allow the existence of a specific dataset such as a catalog to be known, they do not address the issue of locating specific information within that catalog. Such data discovery will typically be supported through the use of a metadata catalog. To integrate metadata access into the NVO, we will define standard metadata access protocols. A candidate for this is the MCAT system developed as part of the Storage Resource Broker (SRB). The MCAT system supports registration of databases as objects in a data handling system. Metadata attributes from the database can be extracted as XML DTDs, and associated data can be aggregated into containers for manipulation within the data grid. This preserves the ability of the data collections to control their metadata, while providing a mechanism to extract both metadata and data for processing within the data grid. Other candidate metadata handling technologies such as CDS GLU have been developed within astronomy. Integrated into Grid security environments, such technology will facilitate the remote query of metadata to permit wide scale data discovery.

Resource and information discovery services will support such functions as *coverage checks* (e.g., based on a uniform sky spatial index, which datasets contain elements near a specified region of interest?); *parameterized service descriptions* (e.g., what services supply “infrared” “images” of “galaxies”?); *hierarchical inventories* (e.g., archive/dataset/service); and *distributed object* interface definitions.

Data Access Services

NVO data access services can be characterized as “collection-specific” and “grid-specific”. Collection-specific services are most often constructed with simple data retrieval and off-the-shelf DBMS functionality, and include catalog subsetting, image metadata subsetting, image archive retrieval, and catalog-reference data (spectra, light-curves, etc.). Most of the existing NVO data collections support direct collection-specific access services. Grid-specific access services rely upon the ability to extract collection objects into containers for manipulation within the data grid, and are similar in terms of the data produced, but may involve substantial computational resources and consequently may require different interaction paradigms as well. Examples include custom science-grade image mosaics (with resampling, reprojection, etc.), data “collections” for specific purposes (images, tables, spectra, and documents all wrapped together and interrelated based on a consumer- or supplier-specific need), and data-based sky background estimates.

Data Access Mechanisms

The underlying storage systems used to hold astronomical data within NVO will be very varied. The astronomical data will be organized into data collections, e.g., for a specific astronomical survey, and within each there will be varied and idiosyncratic content. Each data collection will represent data differently, and NVO cannot realistically hope to impose standards here. Instead, the NVO will develop uniform data and metadata access mechanisms that the collective layer services will use. Interactions with the data collections will be through the connectivity mediators developed for each collection. The NVO will also develop a standard data model on which standard analysis services can be created. A mapping from the data models used by each data collection to the standard model will be developed for each collection. Uniform remote access methods will allow data to be efficiently extracted from any astronomical data collection. These methods need to address both remote access to specific datasets and the movement of large amounts of data between storage systems, e.g., for replication, or pre-staging for computation. We propose to use the GridFTP protocol (and corresponding GridFTP servers) as the basic data movement infrastructure. Of specific interest to NVO will be the development of very high performance parallel (striped) GridFTP servers for bulk data transfer over high-speed networks.

Users of NVO must be able to correctly interpret the complex and heterogeneous data they find. Although astronomical standards like FITS help address this problem, actual data objects are far more complex than the data models (simple images and tables) defined by FITS. The NVO cannot impose standards on how data is stored in archives nor on how data is represented in applications development frameworks. The NVO *will* require standards for metadata and data models within the NVO framework to permit multi-way data translation and construction of NVO services that operate upon diverse data.

The data access architecture of NVO adheres to the usual “hour-glass” nature of Grid infrastructure. At the top, many client toolkits and protocols share a common information model (the narrow part of the hour glass), fanning out at the bottom to interface to many external archives, access protocols, and data representations. The NVO data access layer will provide uniform access to many individual data collections. A major part of our effort will be defining an extensible information model for astronomical data access, encompassing requests, response packaging, and data and metadata handling and queries, including internal NVO standards for astronomical metadata and data models. Data access will include support for data subsetting and filtering as well as format conversion. We will support multiple user-level access protocols (e.g., a simple URL/file-based protocol, XML-based protocols, SOAP, Javaspaces, etc.). This will enable us to target different types of usage while facilitating integration of new access technologies. Each access protocol may in turn support any number of client toolkits (APIs or SDKs) appropriate to different user communities, including support for legacy software frameworks and applications. We will develop a representative set of access protocols and client toolkits to permit end-to-end testing and early user access to multi-wavelength data analysis using NVO.

4.4 The Connectivity Layer

The provider to NVO of a new data collection must publish both the data collection itself and any data-specific *mediators* required to provide uniform access to the data within NVO. Each new collection must also define the

data model used to structure data, and mappings from their collection-specific attributes to the NVO concept space. Mediation is performed by a program module or object wrapper, usually supplied by the data provider, that converts both data and metadata to an NVO-defined standard. The mediator implements standard methods to provide access to the object metadata and data, including any subsetting and filtering of the data. Using a program module for mediation allows access to be optimized for each data collection, regardless of how it is stored, and gives the data provider control over how their data is represented within NVO. For most catalogs, and for the simpler image objects it may be possible to use a generic mediator to access a class of data, with the mediator being driven by an externally defined concept-space for the data. Such a concept-space-driven mediator makes it possible to publish data without writing a program, and allows more scope for optimizing access in different environments (e.g., for large scale supercomputer-based processing).

Data Models

The NVO will be an open system with as low a threshold of participation as possible. Multiple entry points will be geared toward different users, service providers and data suppliers. The principal entry point for data providers is the catalog or data mediator, which maps an external data object into an NVO-defined *data model*. The use of a mediator eliminates requirements on how data is stored in an archive. The mediator maps archival data onto the data models defined by NVO to provide uniform access to data from all branches of astronomy. Data models describe the astronomical objects and classes of objects available through NVO services. More generally, they represent a logical view of astronomical information: objects, collections of objects, ancillary information such as spectral line strengths, and physical models. For example, an “image” object model might contain the following elements: image header, pixel array, variance plane, bad pixel mask, or other masks. The information in the image header is common to both the metadata and the data model. The data model describes all the attributes of an astronomical object. We will investigate the full range of existing astronomical data models and endorse and/or generalize available data models wherever possible. Where no adequate data models exist (e.g., for synoptic data), we will develop new models and test them in collaboration with the suppliers of such data. Our team has the expertise and broad representation of the astronomical community necessary to define such data models (including both the major data providers and the major analysis software providers). This ensures provider buy-in to specific models through their own example, and encourages NVO data models that enable rather than confine.

4.5 The Fabric Layer

Computational Services and Server-Side Data Analysis

The power of NVO derives not just from the federation of various astronomical data collections, but also from the integration of such data into an environment that facilitates discovery through coordinated analysis of data from all branches of astronomy. While coordination of federated collections is handled in a higher layer of the NVO architecture, the resource layer must provide basic mechanisms for allocating and controlling computations on a range of hardware platforms. These mechanisms include software repositories, data staging mechanisms, computational resource management protocols and services, and methods for monitoring and controlling computation. These basic services are provided as part of the Globus toolkit.

If a portion of a calculation can be performed on a single data object it may be best to do the calculation in the server as part of the data access. A sequence of computations may be applied to a given dataset. In doing so, we can optimize access while distributing computation. An NVO research challenge will be to build general mechanisms for constructing enhanced data access services that incorporate local computation. Related efforts include DataCutter (UMd) and the virtual data handling facilities proposed for GriPhyN.

Functions to be applied to astronomical data will normally be specific to a particular class of data, e.g., an image or a spectrum. We will advocate maintaining standard libraries of analysis functions for each class of data. We will research information discovery techniques to allow users to find functions of interest, to publish new functions, and to dynamically download custom functions at access time and apply them to virtual datasets generated in response to the user query.

In distributed systems such as NVO, is it often most cost effective to perform analysis close to where the data resides. Some queries will require joining catalogs containing billions of objects each with hundreds of attributes, and rely on using HTM or other K-D tree techniques to index large catalogs. Reanalysis of image data at the pixel level may be required to dynamically extend catalogs and to refine queries for candidate objects. Remote

computational support for visualization of multi-parametric data may be required in support of cluster analysis. We will investigate the optimal way to manage these remote analyses.

Catalog and Archive Services

Data suppliers span the entire astronomical enterprise, and most current large data suppliers are represented on our team. The largest volumes of data are expected to come from: *service-oriented survey archives*—archives generated by a focused survey, with extensive service to the astronomical community (e.g., 2MASS, SDSS); *service-oriented general archives*—archives of multi-program (and/or multi-instrument) facilities, with extensive service to the astronomical community (e.g., HST, IRSA, CXC); *unserved survey archives*—archives of focused surveys that were performed for specific purposes; and data in the public domain but with no easy access and no service (e.g., MACHO, DPOSS).

Many data suppliers will come to the NVO with sophisticated data models and services already in place (which the NVO will welcome and exploit); they will not need to support new data models to meet their internal needs. They will continue to provide direct server-side object interfaces to users, but will be encouraged to package their data according to protocols endorsed by the NVO. This relatively modest task will involve only server side development with minimal impact on the data provider's architecture.

There are significant differences between the data contents and services of extant archives, and these providers will necessarily drive the initial development of data models. This collaboration will be structured differently for work with the major service-oriented archive centers (e.g., HEASARC, MAST, and IRSA) and for work with the unserved archives. The NVO proposal team includes participation from the major US archive centers, and from providers of some of the largest unserved archives in the world (MACHO, DPOSS, LOTIS, LONEOS).

5 Implementation Plan

5.1 Science Prototypes

The design of the NVO framework will be science driven and ultimately validated by scientific applications. For this reason, a suite of well-defined *science prototypes* will be chosen which will be used to help specify the required NVO system capabilities, and to test and validate those capabilities once they are implemented. A specific test schedule will be developed for each science prototype, with intermediate testing milestones, and a final end-to-end test. The test and validation process will be well-documented, both in terms of the functionality required by the science prototype, and the actual performance of the system as determined by the testing.

5.2 Testbed

The *NVO Testbed* will be created so that, as tools and standards are developed and adapted and initial science applications prototyped, they can be tested in realistic environments with real astronomy data. The testbed will include data, computing, and visualization resources (at Caltech, IPAC, UCSD, selected NPACI and Alliance sites, JHU, STScI, MAST, SDSS), will be accessible to developers and users, and will provide increasingly broad functionality to the user community. *NVO Portals* will provide direct access to large, complementary data sets without requiring additional high-bandwidth connections to the researcher's site. The testbed will incorporate the major data centers and services participating in the NVO federation (CfA/CXC, GSFC/HEASARC, NOAO, NRAO, ADS, ADC). Other data providers (e.g., university-based research groups) will be able to join by implementing simple NVO-compliant interfaces to their data services. Multiple entry portals into the NVO will serve the diverse, discipline-oriented elements of the user community: radio, optical/UV, IR, high-energy, literature and bibliographic services, and education and outreach oriented services. The NVO testbed will be an open system with as low a threshold for participation as possible, and with multiple entry points geared toward different users.

An exciting feature of the NVO testbed will be its inclusion of the major computing resources of the NSF PACI efforts. The PACI sites already have extremely powerful computers and substantial archival storage resources (e.g., the SDSC has an HPSS archive with a capacity of 500 terabytes of storage). The NSF's Distributed Terascale Facility (DTF) program, to be awarded this year, will supplement the PACI resources and services with a much greater emphasis on providing access to large scientific data collections, through very large data caches, expanded archival storage activities, distributed visualization, ultra high-speed network connections among selected sites, and vigorous deployment of grid software. The NVO testbed expects to work closely with the recipients of the DTF

award and provide synergy in this bold step towards building a broader scientific computing environment. The DTF component of the NVO testbed will enable astronomy data investigations that are currently out of reach.

The key activities of the NVO testbed effort will be to:

- Install and test the software adopted or developed by this project, being careful to coordinate versions among all sites participating in the testbed
- Arrange for resource allocation and scheduling procedures on the major computing and data storage resources;
- Develop documentation and provide user support for experimental use of the NVO testbed
- Report problems and, based on user experiences, suggest improvements, to the software/interface developers
- Facilitate the movement of large data collections among the data caches and archival storage resources in the participating sites

In short, the NVO testbed task will provide to this project and to the user community a rich environment on which to test systematically the evolving NVO software and interfaces, with powerful enough resources to obtain new scientific results.

5.3 Implementation Approach and Milestones

We will emphasize the development of an open implementation, enabling the creation of an NVO that links existing and future data archives, query and analysis tools, and human interfaces. The implementation strategy is one of incremental progress, with new capabilities made available to the user community on a regular schedule. We envision the first implementation of a framework usable by a subset of the community to be available within the first year of this project. A detailed development timeline is shown in the Supplementary Documents (Appendix B).

Six *Development Task Teams* including both IT and astronomy expertise will be responsible for individual work packages. They are organized by technical area: (1) *Portals and Workbenches*, (2) *Metadata Standards*, (3) *Grid Services and Testbed*, (4) *Data Models*, (5) *Resource Layer and Data Access*, and (6) *Data and Services*. The correspondence between the teams and the NVO architecture elements is given in the WBS table in the management section. A *Technical Working Group* provides the primary mechanism for communication and coordination between task teams, and a *Science Working Group* is responsible for defining, developing, deploying, and testing science prototypes.

This project will be extending, adapting, and integrating a number of IT tools and techniques so that they can be used to access and analyze many large, existing, and growing astronomy data archives. These archives use different database systems, data layouts, etc., and this heterogeneity will never disappear. One of the biggest challenges is to devise and define *de facto* interface standards that will be effective for federating these databases without requiring undue efforts to conform to the *de facto* standards. User experience with the early NVO prototypes will also be an essential input to guide the development of tools, portals, and workbenches. Therefore, at the highest level, our approach will be to define and implement four major versions of the NVO prototype environment, each to be completed at the end of each of the first four years of the project. This strategy will allow experience with real users and with real grid computing environments to refine successfully our designs. In addition, during the life of the project grid software will evolve and become more capable and robust, thus enabling more sophisticated distributed operations. Furthermore, as the NVO user community gains experience with the early NVO prototypes, usage models and expectations will change. The frequent major releases of the software environment will permit the feedback loop to be effective.

With properly designed interfaces, it will be possible for anyone in the community to add analysis tools and archives: an NVO that can grow gracefully in data and functionality is one of the major design goals. We hope that our framework will actively encourage independent groups to build their own tools and components, and that there will be Announcements of Opportunity by the funding agencies to support such activities. 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 case must be made for each tool, but they should also be general enough that the entire community can use them for research.

6 Outreach and Education

NVO's grand goal to define the future of astronomical research is matched by an equally grand outreach goal. Building on our and our partners' extensive expertise we will integrate into the NVO infrastructure a far-reaching Education and Public Outreach (EPO) program that is innovative, exciting, and effective. There will be four complementary strands, each targeted at a different audience. We will communicate information technology advances as we build the NVO, and astronomy discoveries as the content and scope of NVO grows. We will partner with existing experts and centers of excellence, extending the reach of NVO to a large fraction of the people working in astronomy and technical education and outreach. We will engage student interns in outreach and in building the NVO itself, directly involving the next generation researchers in our work, and provide a wealth of opportunities to bring the benefits of NVO's advances to society as a whole.

For formal education (through college level), we propose to use NVO data as a basis for creating inquiry-based online and hardcopy resources that are directly relevant to curricula. With our partners we will create online resources, together with hard-copy student workbooks and lesson guides to support a wealth of applications. Educational materials will be tied to national science and technology education standards. They will be robustly tested and evaluated using methodology like that developed in support of the STScI "Amazing Space" modules, in particular incorporating independent evaluation. Materials will be made available to mathematics, science and information technology teachers nationwide, using well established national distribution channels, which have been developed by us, our partners, and by NASA. We intend to create interfaces and tools that are student friendly and pedagogically sound. They will be designed to ensure that access to NVO data is intuitive for all users. There will be workshops at conferences such as NSTA and NCTM to create "master" teachers who are intimately familiar with the use of these resources and tools. Through such workshops, and with partners like Gettysburg College's Project CLEA (Contemporary Laboratory Experiences in Astronomy) and U.C. Berkeley's CSE@SSL (Center for Science Education at the Space Sciences Lab), we will use existing teacher professional development opportunities to train educators in effectively using NVO. In the process, we'll develop "NVO Ambassadors" who take their expertise back to their school districts to share with their colleagues for maximum leverage. We will work with community college teachers to create tools and specific online products tailored to the needs of this presently underserved user community. We will create undergraduate and graduate internships in information science and NVO-based astronomy. We will give all interns experience in developing and using the educational and outreach component of NVO, while they help develop the NVO infrastructure.

Selected Partners Committed to NVO Outreach

- Association of Science-Technology Centers
- International Planetarium Society
- National Air and Space Museum
- Silicon Graphics (Digital Planetarium)
- Spitz (Electric Sky)
- Maryland Space Grant Consortium
- Gettysburg College (Project CLEA)
- U.C. Berkeley (CSE@SSL)
- American Museum of Natural History

For informal education, NVO will be integrated with museums, science centers, and planetariums. Collaborating with professional organizations such as the Association of Science-Technology Centers, the Museum Computer Network, and the International Planetarium Society, we will develop appropriate NVO interfaces and tools, publicize their existence, and promote their adoption by content developers. With our partners we will work to ensure that the access tools match pedagogical needs, are simple to use, and have meanings that are transparent to the users, be they students, teachers or interested members of the public. NVO data will be readily usable by Informal Science educators, exhibit designers, and program developers, and we will create tools for converting NVO data into forms readable by common bitmap editing programs, such as Adobe Photoshop, and into forms readable by common interactive multimedia production tools like Macromedia Director. We propose creation of tools to export NVO data to digital all-dome theater systems, such as Silicon Graphic's *Digital Planetarium*, Evans and Sutherland's *Star Rider*, Spitz's *Electric Sky*, and Sky-Skan's *SkyVision*. We will catalyze the creation and dissemination of replicable end-user products for inquiry-based exploration of NVO data sets by the general public in informal settings, and partner with leading informal science institutions to develop and widely distribute interactive "kiosk" applications that teach astronomical concepts and allow unguided exploration of selected sets of NVO data. We will help them develop and distribute "kits" of live demonstration scripts that teach astronomical concepts using NVO data, and enable leading planetariums to develop and widely distribute panoramic, and all-dome digital representations of key astronomical concepts. NVO's science value will be highlighted with astronomical discoveries made using archived data. We will partner with leading planetariums to develop and

widely distribute planetarium programs about serendipity in astronomy. We will partner with an established museum exhibit developer on a major touring exhibit highlighting the discoveries made possible by NVO, and the modern technology enabling NVO.

For online outreach, we aim to become *the* portal for public and education community access to astronomy. There will be a public/educator portal to the NVO, an “Encyclopedia Galactica”, and an intuitive user-friendly interface to the NVO for novice and advanced users. We will provide rich content describing NVO, how it works, background information and “guided tours” of interesting objects. Online access to NVO EPO products will range from lesson plans and curriculum products to press releases, multimedia products, and educational games (e.g., nvo@home, spot the high redshift Supernova, find the asteroid or Kuiper Belt Object). We will provide a portal for amateur astronomers to use the NVO, and then submit or share their own data with other NVO users

For the news media, we will create an easy-to-access online resource of NVO-derived astronomical images specifically for the news media. We capitalize on the close links with the media established by the Hubble program to bring late breaking discoveries made by NVO to the attention of news media and science journalists. We will create a mechanism whereby investigators using NVO can release their findings to journalists with appropriate online support materials, and create background material describing the science *and* the information technology used, and place them in a news archive section of the portal. The tremendous public interest in space science and astronomy will also provide an excellent opportunity to teach the public about information technology.

The EPO activities will be phased through the duration of the project. In years 1 and 2 the main focus will be on creating the web portal and establishing the infrastructure. The primary EPO program occurs in years 3–5, as NVO itself demonstrates its capabilities.

7 Relationship to Other Projects

The NVO framework will be constructed from a combination of pre-existing components (COTS, public domain packages, SRB, Globus, etc), adaptations thereof, and newly developed tools and systems. We will draw heavily upon the groundbreaking work of the GriPhyN project [GRIP], already supported by NSF’s ITR program, and on existent or emerging technologies and standards such as XML and SOAP-based web services. The NVO framework will also incorporate subsystems developed through other NSF ITR programs (several specific programs are noted in the Budget Narrative) and complementary programs run by other agencies. The US NVO effort is also being coordinated with international VO development activities, particularly in the areas of metadata standards and interoperability but also in a general exchange of information concerning technical implementation strategies, in order to assure that the end result is a Global Virtual Observatory. The NVO initiative will return tools and components to the IT and scientific communities for re-use and adaptation in other distributed information management services.

The NVO has very close ties to the GriPhyN and iVDGL projects. The NVO’s current data holdings are comparable in size to today’s high-energy physics experiments, although the data has a much less hierarchical structure than the HEP data. GriPhyN’s primary focus is Virtual Data. The iVDGL is building a concrete, large-scale testbed for grid-based data analysis and computing, and the NVO will begin by federating existing archives and metadata standards. In the long run there will be an inevitable convergence of these approaches. We believe that within the time frame of the current proposals such a convergence will be achieved: the Grid will become a reality, and the NVO will be ready to use its technology and resources. It would be extremely wasteful to duplicate efforts. The Virtual Data-Grid toolkit is designed by GriPhyN, it has essentially settled the issues of which standards to adopt for such services, and the NVO has decided to adopt and use those. The two groups have been collaborating for several years: Szalay, Newman, and Gray collaborate on a joint NSF KDI project, and Szalay, Kent, R. Moore, Williams, Foster, Livny, and Kesselman are participants on GriPhyN, iVDGL, and the NVO. In particular, Szalay, Kesselman and Livny will serve as liaisons among the projects. This close collaboration will not lead to “double-dipping,” but rather will create a well-defined coordination among these efforts.

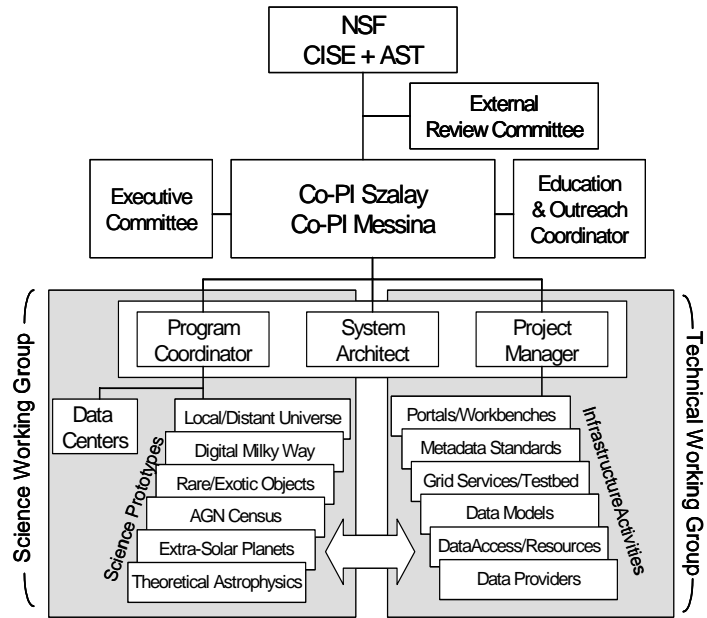
We are aware of several complementary medium-sized proposals that have been submitted to the ITR. These include *Common grid-Based Radio Archive (COBRA)* (T. Cornwell, NRAO, PI), *Developing the National Virtual Observatory Data Model* (A. Goodman, Harvard, PI), and *Statistical Data Mining for Cosmology* (A. Moore, CMU, PI). Given the excitement of the community about research in this area we expect there are several more small or medium-size proposals unknown to us.

8 Management Plan

The challenge of building a framework to enable the NVO will be met with a management structure that supports distributed research and development. We join distributed, multi-disciplinary, heterogeneous resources with strong communication and coordination. We take optimal advantage of the domain expertise already resident in the organizations supporting the existing archival systems, sky surveys, and source catalogs of the astronomy community and meld this diversity with state-of-the-art information technology. Our structure ensures accountability to both the community and the funding agency and ensures that astronomy needs drive technology development.

Our team is led by co-PIs Dr. Alexander Szalay at The Johns Hopkins University (JHU) and Dr. Paul Messina at California Institute of Technology (CIT). The contractual relationship with NSF is via JHU, overseen by Szalay, an astronomer with a strong CS background who has pioneered applying advanced computing techniques to research with large astronomy databases. Szalay holds appointments in both the Physics and Astronomy and Computer Science Departments at JHU. He is a key member of the Sloan Digital Sky Survey project, designing its archive, and a prime mover in the National Academy report that strongly recommended an NVO. He will ensure the project is driven by the needs of astronomy. Messina is a computational scientist who has formed and led a number of multidisciplinary research projects, including the Scalable I/O Initiative, the CASA gigabit testbed, and the Concurrent Supercomputing Consortium. He is Chief Architect of the National Partnership for Advanced Computational Infrastructure and has led DOE's Accelerated Strategic Computing Initiative. Messina oversees the essential IT developments in federating large, heterogeneous databases, enabling efficient searches and cross-correlations in TB+-scale databases, and enabling the exchange of metadata across diverse systems.

Szalay and Messina co-chair an *NVO Executive Committee*, responsible for approving changes to the project's technical development plan, reviewing progress, and re-allocating funds among tasks as necessary. In the core management group, the co-PIs are augmented by a Program Coordinator (PC) and a Project Manager (PM). The PC assists the PIs in ensuring that development tasks meet high-level science goals and that incremental capabilities are delivered to the end-user/astronomy community. The PM monitors schedules, budgets and progress of the development task teams. A System Architect (SA) will be appointed to help coordinate the overall system engineering and design function, including interfaces between the architectural elements. Six *Development Task Teams* including both IT and astronomy expertise will be responsible for individual work packages. They are organized by technical area: (1) Portals and Workbenches, (2) Metadata Standards, (3) Grid Services and Testbed, (4) Data Models, (5) Resource Layer and Data Access, and (6) Data and Services. The correspondence between the teams and the NVO architecture elements is given in the WBS table that follows. The core management group is responsible for choosing a leader for each Development Task Team, who reports to the PM. A *Technical Working Group* provides the primary mechanism for communication and coordination between task teams. This group is chaired by the PM and includes the leads of the Development Task Teams, the PC, and the SA. A *Science Working Group* is responsible for defining, developing, deploying and testing science prototypes. This group is chaired by the PC and includes application scientists, the PM, and the SA.



An Education & Outreach Coordinator ensures NVO benefits education and the public at large. Designated liaison team members ensure coordination with related international efforts. Periodic reviews by an *External Oversight Committee* assure NSF that the NVO framework is responsive to both astronomy and IT community needs.

8.1 Project Structure

We have developed a standard work breakdown structure (WBS) to organize the management, development, test-bed deployment, and education and outreach elements of the project. The level of effort (full-time equivalents, or FTE) in each WBS area has been profiled for the five-year duration of the project, focusing first on core technologies and prototype development and later on system integration and deployment. The project WBS and its correspondence to the NVO architecture elements are shown below in both tabular and graphical form.

WBS	Activity	Architecture Elements	Year					Total
			1	2	3	4	5	
1	Project Management		2.40	2.40	2.40	2.40	2.40	12.00
2	Systems Architecture	all	1.00	1.00	0.50	0.50	0.50	3.50
3	Portals and Workbenches	A	0.75	1.50	2.50	2.50	2.00	9.25
4	Metadata Standards	B, C, E	4.10	3.05	1.05	0.50	0.50	9.20
5	Grid Services	D	1.20	2.20	2.20	0.50	0.50	6.60
6	Data Models	F	3.75	3.75	2.15	0.95	0.50	11.10
7	Data Access/Resource Layer	E, F	3.25	3.25	2.25	1.25	0.60	10.60
8	Service/Data Provider I&I	F, G	0.95	2.15	3.40	3.40	1.80	11.70
9	Test-Bed	all	0.75	0.75	1.75	2.75	3.15	9.15
10	Science Prototypes	all	2.00	2.00	3.00	3.00	4.00	14.00
11	Outreach and Education		1.00	1.00	1.00	4.00	5.00	13.00
	Totals		21.15	23.05	22.20	21.75	20.95	110.10

The contributions of each organization to the work-breakdown structure have been calculated based on each organization's expertise. The existing astronomical data centers contribute most heavily in WBS areas 3, 4, 6, 7, and 8, while the IT/CS organizations focus most strongly in areas 2, 4, 5, and 9. There is substantial overlap, of course, in order to assure that domain expertise—be it astronomical or IT—is taken into account in all aspects of system design and implementation. The budget narrative provides more detail on the specific levels of effort of the collaborators. The table below highlights the interest and expertise areas of each organization.

Communications will be extremely important, and frequent group telecons and ~quarterly group meetings will be an essential component of the development environment. Collaborative distributed development tools will be utilized as necessary to assure code integrity. Project-wide software engineering standards will be defined, based on previous experience with distributed development efforts, and enforced.

8.2 Project Management

We have successfully met the difficult challenge of assembling a proposal team to begin work on the infrastructure for the NVO, with representation from all major astronomical data centers and service providers partnered with established leaders in the IT/CS community. We recognize fully the forthcoming challenge of actually managing a project of this complexity and scope. There are 16 funded organizations (19 when one realizes that there are actually four autonomous groups at Caltech), two unfunded collaborative organizations, external education/outreach collaborators and international collaborators. Our project management budget is modest at ~12.5% of the overall staffing, not counting the management efforts of the team and group leads who will have responsibility for design and implementation of specific work packages. Fortunately our team includes senior personnel with prior experience and proven results in managing large, physically distributed projects, including substantial international partnerships. The NVO is by its nature a distributed system, and effective management of distributed resources applies equally to the data systems and services and the development and implementation team.

Our project management approach will rely upon clearly defined work packages, with aggressive but achievable schedules and agreed-upon deliverables. Participating organizations will be held accountable for their contracted work packages, and the project executive committee will have the authority to curtail or terminate participation of groups who do not produce. Work packages will have terms no longer than one year in duration with progress check-points at least quarterly, allowing problem areas to be identified early and resolved without waste of resources.

<u>Organization</u>	<u>Areas of Expertise</u>
Caltech	
Astronomy Department	Large surveys, science drivers, database federation
IRSA/IPAC	Data provider services, metadata standards, portals and workbenches, large surveys
NED	Metadata standards, data provider services
CACR	Systems architecture, NVO services, testbed, project management
JHU	Science oversight and testing, database federation, high performance database engine
STScI	Metadata standards, data provider services, project management, outreach/education
ADC	Metadata standards, metadata delivery
Argonne/U Chicago*	Grid technologies
CXC/Harvard	Data models, data provider services
FNAL	Large surveys, database federation, testbed
HEASARC	Metadata standards, data provider services, portals and workbenches
LLNL*	Large surveys, metadata standards, science drivers
Microsoft Research	Large high performance databases, database federation
NCSA/UIUC	Metadata standards, metadata delivery, data discovery
NOAO	Data delivery and data mediators, large surveys, science drivers (theory)
NRAO	Metadata standards, data delivery and data mediators
SDSC/UCSD	System architecture, NVO services, data access
U Penn	Large surveys, metadata standards, data models, science drivers
U Pitt/CMU	Knowledge management, analysis, visualization
U Wis	NVO services
USC	NVO services, grid technologies
USNO	Data provider services, metadata standards

* Unfunded collaborator

9 Results from Prior NSF Support (A. Szalay)

AST-9802980, “The Structure of the Universe Beyond 100 Mpc,” is supporting research on the large-scale distribution of galaxies in the Universe. The goals of the proposal are to develop innovative statistical techniques that enable optimal signal-to-noise determinations of the power spectrum of the density fluctuations in the Universe. There are a few months left in the grant, and most of the planned research has successfully been accomplished. We have designed a novel photometric redshift technique, and successfully applied it to studies of the evolution of galaxy clustering and star-formation rates. We have built an analysis package to measure the redshift-space power spectrum of the galaxies based upon the Karhunen-Loeve transform, generalized previous approaches to an arbitrary survey geometry, and successfully tackled various computational issues. As a practical test, we have applied the technique to the LCRS redshift survey, and provided the best constraint at the time on the redshift distortion parameter and the small-scale motion of the field galaxies. The technique has the promise to be able to measure the cosmological constant to an accuracy of 2%. This work has resulted in 19 publications to date.

The grant AST-SGER-9876645, “Towards a Virtual Observatory,” has supported Szalay’s activities in identifying key issues related to the NVO. It has enabled him to travel to various conferences, do various pilot studies and partially sponsor meetings at JHU where the ideas on how to build the NVO started to crystallize. This led to over 10 publications.

The grant KDI-9980044, “Accessing Large Distributed Archives,” is a collaboration with Caltech and Fermilab and deals with issues of distributed database access and efficient data clustering techniques, which enable us to analyze large data sets efficiently. This has led to over 10 publications.

References

- [2MASS] *The Two Micron All Sky Survey at IPAC*, <http://www.ipac.caltech.edu/2mass/>
- [ACC99] *Describing Astronomical Catalogs and Query Results in XML*, A. Accomazzi et al., <http://vizier.u-strasbg.fr/doc/astrores.htm>
- [BL01] *The Semantic Web*, T. Berners-Lee, J. Hendler & O. Lassila, in *Scientific American*, 284, 5 (May 2001) p. 34.
- [CON01] *Fast Algorithms and Efficient Statistics: Density Estimation in Large Astronomical Datasets*, A.J. Connolly, C. Genovese, A.W. Moore, R.C. Nichol, J. Schneider, L. Wasserman, 2000, AJ, in press
- [DE] Digital Earth program: <http://www.digitalearth.gov/>
- [DPOSS] *Digital Palomar All-Sky Survey*, S.G. Djorgovski, S. Odewahn, R. Brunner, R. Gal, A. Mahabal, <http://www.astro.caltech.edu/~george/dposs/>
- [DS00] *The Digital Sky Project: Federating the Multi-Wavelength Sky Surveys*, <http://www.digital-sky.org/>
- [EUUS] European Union - United States joint workshop on Large Scientific Databases, funded by the NSF and EU, R. Williams, P. Messina, F. Gagliardi, J. Darlington, G. Aloisio, <http://www.cacr.caltech.edu/euus/>
- [FK99] *The Grid: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, eds., Morgan Kaufmann, 1999
- [GRIP] *GriPhyN: Grid Physics Net*, a consortium of computer science activities to service physics experiments, <http://www.griphyn.org/>
- [GUI98] *Astronomical Markup Language*, D. Guillaume, <http://monet.astro.uiuc.edu/dguillau/these/>
- [HAN00] *ISAIA: Interoperable Systems for Archival Information Access*, R. Hanisch, T. McGlynn, R. Plante, J. King, R. White, J. Mazarella, <http://heasarc.gsfc.nasa.gov/isaia>.
- [IRSA] The archive node for NASA's infrared and submillimeter astronomy projects and missions, <http://irsa.ipac.caltech.edu/>
- [KST01] *The Hierarchical Triangular Mesh*, P.Z. Kunszt, A. S. Szalay, and A. R. Thakar, Proc. Mining the Sky, A. Bandy, ed., Kluwer, 2001
- [NAS99] *Astronomy and Astrophysics in the New Millennium (Decadal Survey)*, National Academy of Science, Astronomy and Astrophysics Survey Committee, <http://www.nap.edu/books/0309070317/html/>
- [NED] NASA Extragalactic Database, JPL/Caltech, <http://ned.ipac.caltech.edu>
- [NVO] *Virtual Observatories of the Future*, R. J. Brunner, S. G. Djorgovski, & A. Szalay, eds., Astron. Soc. Pac. (to be published), proc. of the conf Virtual Observatories of the Future, Caltech, June 2000. Presentations at <http://www.astro.caltech.edu/nvoconf/presentations.htm>. See also *Towards a National Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan* (white paper) http://www.astro.caltech.edu/nvoconf/white_paper.pdf
- [ORBIT] *An Interface to Astronomical On-line Archive Science Information Services* <http://irsadev.ipac.caltech.edu:8002/Applications/Oasis/>
- [RBP99] *Cross-Association of ROSAT/Bright Source Catalog Sources with the USNO A2 Optical Point Sources*, R.E. Rutledge, R.J. Brunner, T.A. Prince, C. Lonsdale, submitted to ApJ (1999)
- [SB99] *Astronomical Archives of the Future, A Virtual Observatory*, A. Szalay and R. Brunner, Fut. Gen. Comput. Sys. 16 (1999) 63
- [SDSS] *The Sloan Digital Sky Survey*, A.S. Szalay, in Computing in Science and Engineering, V1-N2, p.54, 1999
- [SDSS2] *The Indexing of the SDSS Science Archive*, P.Z. Kunszt, A.S. Szalay, I. Csabai, and A. Thakar, in Proc ADASS IX, eds. N. Manset, C. Veillet, D. Crabtree, (ASP Conference series), 216, 141 (2000)
- [SKYSRV] *Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey*, A.S. Szalay, P. Kunszt, A. Thakar, J. Gray, D. Slutz, and R. Brunner, R., Proc. SIGMOD 2000 Conference, 451, 2000
- [SRB] *Storage Resource Broker*, San Diego Supercomputer Center, <http://www.npaci.edu/DICE/SRB/>
- [SZ01] *Prototype data discovery with AI tools*, A. Szalay, private communication, 2001
- [UDDI] *Universal Description, Discovery, and Integration*, <http://www.uddi.org/>

- [VS] *VirtualSky: A High-Resolution Sky Survey Online*, R. Williams, A. Szalay, J. Gray, S.G.Djorgovski, J. Bunn, R. Brunner, <http://virtualsky.org/>
- [WSDL] *Web Services Description Language*, <http://xml.coverpages.org/wsdl.html>
- [XSIL] *Extensible Scientific Interchange Language, an XML Dialect*, R. Williams, <http://www.cacr.caltech.edu/XSIL/>

List of Personnel

Principal Investigators

Paul Messina	California Institute of Technology	Alex Szalay	The Johns Hopkins University
--------------	------------------------------------	-------------	------------------------------

Senior Personnel

Chalres Alcock	University of Pennsylvania	Carol Lonsdale	Caltech / NASA Infrared Processing and Analysis Center
Kirk Borne	NASA Goddard Space Flight Center / Raytheon	Thomas McGlynn	NASA Goddard Space Flight Center / USRA
Tim Cornwell	National Radio Astronomy Observatory	Andrew Moore	Carnegie-Mellon University
David DeYoung	National Optical Astronomy Observatories	Reagan Moore	University of California, San Diego
Giussepina Fabbiano	Smithsonian Astrophysical Observatory	Robert Nichol	Carnegie-Mellon University
Alyssa Goodman	Harvard University	Jeff Pier	United States Naval Observatory, Flagstaff Station
Jim Gray	Microsoft Research	Ray Plante	University of Illinois, Urbana-Champaign
Robert Hanisch	Space Telescope Science Institute	Thomas Prince	California Institute of Technology
George Helou	Caltech / NASA Infrared Processing and Analysis Center	Ethan Schreier	Space Telescope Science Institute / The Johns Hopkins University
Stephen Kent	Fermi National Accelerator Laboratory	Nicholas White	NASA Goddard Space Flight Center
Carl Kesselman	University of Southern California	Roy Williams	California Institute of Technology
Miron Livny	University of Wisconsin, Madison		

Collaborators

Bruce Berriman	Caltech / NASA Infrared Processing and Analysis Center	John Good	Caltech / NASA Infrared Processing and Analysis Center
Roger Brissenden	Smithsonian Astrophysical Observatory	Barry Madore	Caltech / NASA Infrared Processing and Analysis Center
Robert Brunner	California Institute of Technology	Joseph Mazzarella	Caltech / NASA Infrared Processing and Analysis Center
Cynthia Cheung	NASA Goddard Space Flight Center	Brian McLean	Space Telescope Science Institute
Kem Cook	Lawrence Livermore National Laboratory	Marc Postman	Space Telescope Science Institute
Andrew Connolly	University of Pittsburgh	Arnold Rots	Smithsonian Astrophysical Observatory
David Curkendall	NASA Jet Propulsion Laboratory	Steven Strom	National Optical Astronomy Observatories
George Djorgovski	California Institute of Technology	Anirudha Thakar	The Johns Hopkins University
Ian Foster	University of Chicago	Douglas Tody	National Optical Astronomy Observatories
Roy Gal	The Johns Hopkins University		

International Liaison

Piero Benvenuti	European Space Agency	Fionn Murtagh	University of Belfast
Daniel Durand	Canadian Astronomy Data Centre	Ray Norris	Australia Telescope National Facility
Francoise Genova	Centre de Donnees astronomiques de Strasbourg	Sadanori Okamura	University of Toyko
Andrew Lawrence	Royal Observatory Edinburgh	Peter Quinn	European Southern Observatory

Appendix A: Major U.S. Astronomy Data Holdings

This table is the source of the data holdings chart shown on page one of the main proposal. Information is coded as follows:

Name: Name of mission, facility, survey, or data set

Status: L = legacy data, A = active mission or facility with increasing data

Origin: SM = space mission, GB = ground-based facility, GBS = ground-based survey

Data Type: S = spectral, I = image, P = photometry (time series, polarimetry), C = catalog, R = raw (unprocessed), B = bibliographic

Spectral Coverage: Wavelength, frequency, or energy bands included in the data set

Size: Volume of data

Number of Observations: Number of spectra, images, catalogs, etc.

Number of Sources: Number of observed targets, or number of extracted objects

Responsible Organization: Curator of the data set

Name	Status	Origin	Data Type	Spectral Coverage	Size	Number of Observations	Number of Sources	Responsible Organization
IUE	L	SM	S	1200-3350 Å	600 GB	104,000	10,000	STScI/MAST
EUVE	L	SM	S,I	70-760 Å	61 GB	1150	400	STScI/MAST
Copernicus	L	SM	S	900-3150 Å	1 GB	687,000 scans	551	STScI/MAST
ASTRO UIT	L	SM	I	1200-3300 Å	56 GB	1600	200	STScI/MAST
ASTRO HUT	L	SM	S	825-1850 Å	0.6 GB	500	300	STScI/MAST
ASTRO WUPPE	L	SM	P	1400-3200 Å	0.1 GB	400	200	STScI/MAST
ORFEUS BEFS	L	SM	S	900-1200 Å	4 GB	317	108	STScI/MAST
ORFEUS TUES	L	SM	S	900-1400 Å	0.6 GB	239	62	STScI/MAST
ORFEUS IMAPS	L	SM	S	950-1150 Å	0.3 GB	600	10	STScI/MAST
HST	A	SM	S,I,P	1100-25000 Å	>7 TB	>230,000	>20,000	STScI/MAST
FUSE	A	SM	S	905-1190 Å	>25 GB	>1160	>800	STScI/MAST
DSS	A	GBS	I	4400-7000 Å	465 GB	4780 plates		STScI/MAST
VLA FIRST	A	GBS	I	20 cm	110 GB	15,000	720,000	STScI/MAST
GSC II	A	GBS	I,C	4000-8500 Å	2 TB	3500 plates	2 x 10 ⁹	STScI/MAST
SDSS Early Release	A	GBS	I,S,C	3600-9200 Å	1 TB	15,000	10 ⁷	STScI/FNAL/JHU
SDSS	A	GBS	I,S,C	3600-9200 Å	20 TB	1 x 10 ⁶	3 x 10 ⁸	FNAL/JHU/STScI
Ariel V	L	SM	P,C	0.3-40 keV	0.1 GB	250		GSFC/HEASARC
ASCA	L	SM	S,I,P,C	0.4-10 keV	535 GB	3833		GSFC/HEASARC
BBXRT	L	SM	S,P	0.3-12 keV	2.1 GB	157		GSFC/HEASARC
Beppo-SAX	A	SM	S,I,P	0.1-300 keV	50.3 GB	3462		GSFC/HEASARC
CGRO	L	SM	C	30 keV - 30 GeV	279 GB	~10,000		GSFC/HEASARC
COS-B	L	SM	I	2 keV - 5 GeV	0.1 GB	252		GSFC/HEASARC
Copernicus UCLXE	L	SM	R	0.5-10 keV	0.4 GB	6555		GSFC/HEASARC
Einstein (HEAO-2)	L	SM	S,I,P,C	0.2-4.5 keV	15 GB	~6000		GSFC/HEASARC
EUVE	L	SM	S,I	70-760 Å	61 GB	1150	400	GSFC/HEASARC
EXOSAT	L	SM	S,I,P,C	0.05-20 keV	107 GB	6614		GSFC/HEASARC
Ginga (Astro-C)	L	SM	S,P	1-500 keV	19.7 GB	11,673		GSFC/HEASARC
HEAO-1	L	SM	S,I,P,C	0.2 keV - 10 MeV	9.5 GB	~10,000		GSFC/HEASARC
OSO-8	L	SM	P	0.15 keV - 1 MeV	6.4 GB	~2500		GSFC/HEASARC
ROSAT	L	SM	S,I,P,C	0.1-2.5 keV	181 GB	15,000		GSFC/HEASARC
RXTE	A	SM	S,P,C	2-250 keV	883 GB	24,561		GSFC/HEASARC
SAS-2	L	SM	I	20 MeV - 1 GeV	0.1GB	81		GSFC/HEASARC
SAS-3	L	SM	R	0.1-60 keV	7.2 GB	321		GSFC/HEASARC
Vela 5B	L	SM	P	3-750 keV	0.1 GB	268		GSFC/HEASARC
Uhuru (SAS-1)	L	SM	C	2-20 keV	0.1 GB	339		GSFC/HEASARC
XMM	A	SM	S,I,P,C	.1-15 keV	1.5 GB	3		GSFC/HEASARC
HEASARC Catalogs	A	SM/GBS	C	mostly > 1 keV	2 GB	~1 x 10 ⁶		GSFC/HEASARC
NEAT/SkyMorph	A	GBS	I,C	7,000 A	2.5 TB	80,000	4 x 10 ⁸	JPL-GSFC/HEASARC
WENSS	A	GBS	I	325 MHz	0.5 GB	500		HEASARC/SkyV.
SUMSS	A	GBS	I	843 MHz	1.0 GB	108		HEASARC/SkyV.
GB6	L	GBS	I	4850 MHz	2.5 GB	616		HEASARC/SkyV.
IRAS	L	SM	I,C	12,25,60,100 μ	10 GB	1720	1 x 10 ⁶	IPAC/IRSA
MSX	L	SM	I,C	8.3 μ	200 GB	1590	330,000	IPAC/IRSA
2MASS	A	GBS	I,C	1.25, 1.65, 2.17 μ	15 TB	4 x 10 ⁶	3 x 10 ⁸	IPAC/IRSA
IRAS	L	SM	I,C	12,25,60,100 μ	1720	1720	1 x 10 ⁶	GSFC/NSSDC
COBE	L	SM	S,I,P	1.25 μ - 10 mm	28 GB	3 x 10 ⁹	n/a	GSFC/NSSDC
SWAS	A	SM	S	487-557 GHz	1.1 GB	6000	77	GSFC/NSSDC
Chandra	A	SM	S,I,P,C	0.3-8 keV	1 TB	3000	500	SAO/CXC
Mosaic North	A	GBS	I		10 TB	72,000		NOAO
NDWFS	A	GBS	I,C		300 GB			NOAO
NOAO Surveys	A	GBS	I,C		2 TB			NOAO
NVSS	A	GBS	I,C					NRAO
FIRST	A	GBS	I,C	20 cm	110 GB	15,000	720,000	NRAO
VLA	A	GB	I,R	0.7-400 cm	2.5 TB	20,000		NRAO
VLBA	A	GB	I,R	0.3-90 cm	7.5 TB	3000	~3000	NRAO
GBT	A	GB	S,I,R	0.3-90 cm				NRAO
DPOSS	L	GBS	I,C	4000-8500 Å	3 TB	3500 plates	1 x 10 ⁹	Caltech
USNO-A2.0	L	GBS	I,C	4000-8500 Å	10 TB		5.2 x 10 ⁸	USNO
MACHO	L	GBS	I,P,C	4500-7600 Å	7 TB	93,000	7 x 10 ⁷	UPenn/LLNL
OGLE II	L	GBS	I,P,C	6800-8400 Å	1 TB	30,000	4 x 10 ⁷	Upenn/Princeton
LOTIS	A	GBS	I	4000-8000 Å	12 TB	90,000	2 x 10 ⁷	UPenn/LLNL
LONEOS	A	GBS	I,C	4000-8000 Å	2 TB	60,000	3 x 10 ⁸	UPenn/LLNL
ADC	L		C	various	18 GB	3500 catalogs		GSFC/ADC
NED	A,L	SM/GBS	S,I,P,C,B	various	155 GB	4.2 x 10 ⁶	3.3 x 10 ⁶	IPAC/NED
ADS	A		B		350 GB	>2.2 x 10 ⁶ abstracts & references		SAO

Appendix B. Implementation Timeline and Milestones

Year 1

- Identify what portions of what collections will be available through NVO by end of the year
 - Work with archives to define their content
 - Plan and prototype directory service for NVO
 - Work with archives and users to scope an XML toolkit
- Identify science drivers and their infrastructure requirements
 - Create documents defining each science goal -- framed as IT requirements
 - Build a small team for each science goal with implementation plan
 - Plan data, networking, and caching requirements for each science goal
 - Plan porting of necessary analysis software to NVO grid environment.
- Define and begin forming initial interface conventions
 - Ways to represent data schemas
 - Transport protocols and presentation standards
 - Protocols for catalog search and image retrieval for all major archives
- Portals
 - Identify commonalities between existing portals
 - Define requirements and existing tools for Grid portal
- Implement and deploy the interfaces for at least three data collections and two analysis tools
 - Use new technologies as well as traditional
- Create an initial data storage plan for derived data collections
 - Specify initial secondary and tertiary storage locations
 - Define an initial data migration plan for implementation and deployment
- Identify initial minimal Grid software suite
 - Identify or build cross-compare software, image mosaicing software
 - Plan server-side versions of common astronomical software
 - Deploy and test Grid software and services across JHU, Caltech, NCSA, SDSC, STScI, CfA, and Fermilab
- Outreach
 - Develop new training and content for astronomers, plan training events
 - Run 2-day workshops in Pasadena and Baltimore for local teachers
 - How to use astronomy data for teaching
 - Restructuring of astronomy portals for education
- Move forward on design and establishment of an international information infrastructure for astronomy

Year 2

- Deploy prototype directory of services
 - Refine NVO profile directory
 - Refine NVO data archive directory
 - Solidify XML schema for profile definition
- Deploy Data Access Layer and its tools
- Build translation tools where necessary
- Build XML toolkit for archives and users
 - Find or build FITS-XML dialect, define extensions to FITS
- Build and test Grid portal for cross-match, mosaicing, etc.
- Develop and deploy visualization capability for massive multi-parameter datasets
 - Integration with clustering/outlier software
- Ensure that data discovery and comparison tools are now mature and in routine operation
- Ensure that major computing resources are in use for significant science goals
- Incorporate requirements from the theoretical astrophysics community
- Establish partnerships with international organizations to assure interoperability of US and non-US facilities

Outreach:

- Work with schools and colleges to enhance curriculum materials based on NVO data and services
- Internship program for undergraduates well-established
- NVO web site deployed, with links to other major astronomy outreach web sites
- NVO web provides data services for other educational activities
- Development of interactive kiosk for museums

Year 3

- Refine interface definitions based on experience with NVO prototype 2
- Carry out several large science runs
- Do terascale runs of the multiparameter analysis package
 - Scientifically significant cluster/outlier search
- Identify scaling issues for software and other that will need to be addressed for the eventual production NVO
- Data access layer deployed and in routine operation
- Deployment of NVO data-publishing portal
- Deployment of NVO Grid portal
- Establish support for higher-level data products, such as pre-prepared cross-identifications
- Outreach
 - Full deployment of NVO portal
 - Interactive kiosk deployed in at least two science museums
 - Undergrad and graduate intern programs

Years 4 and 5

- Conduct large science runs
- Carry out large-scale comparisons of data with simulations
- Rebuild implementation software around converged standards
- Begin deployment of the production NVO
- Provide broad outreach
 - Astronomy as a vehicle for IT education in US schools
- Demonstrate achievement of major goals of this proposal
 - Comprehensive semantic web structure for astronomical data
 - Terascale data and terascale computing for astronomers
 - Diverse, yet interoperable toolkits available to disseminate astronomical data to all who want it
- International consensus on astronomy infrastructure

Budget Narrative

The proposal requests funds for five years, commencing 1 October 2001, to address expenses at 16 organizations. This narrative covers all 16 organizations, identifying specific tasks and responsibilities in synopsis form.

Personnel: The bulk of the budget request is for labor costs at the collaborating organizations. Each organization has used internally approved inflation rates (~3-4% per annum). See below for additional information on the labor allocations for each organization.

Fringe Benefits: Each organization assesses fringe benefits on their base labor costs, and rates vary in the range ~25-30%.

Equipment: Each organization has an equipment budget not to exceed \$10k per year in up to two of the five years. This funding level is adequate to provide workstations or small workgroup servers. Major computational facilities will be provided through institutional partnerships (Fermilab, NCSA, San Diego and Pittsburgh supercomputer centers, NPACI) and through existing computational infrastructure at each organization. [Note: the Caltech budget represents four organizations: the Caltech Astronomy Department, Infrared Science Archive Center, NASA Extragalactic Database, and Center for Advanced Computational Research.]

Travel: Each organization has a travel budget not to exceed \$5k per year per FTE to allow participation in team and working group meetings, collaborative development efforts, and attendance at relevant conferences and workshops. In addition, the JHU budget carries a line item supporting general proposal team meeting costs (conference rooms, A/V equipment rental).

Materials and Supplies: Each organization has budgeted as required for incidental materials and supplies. JHU carries a line item to cover costs for team and workgroup conference calls.

Publications: A small budget is allocated for publications of papers related to the project in peer-reviewed journals.

Computer Services: Each organization has budgeted as required for computer services (network connection fees, maintenance expenses, etc.) Direct charges for computer services are often a form of institutional overhead and are typically based on the number of funded personnel.

Other: Each organization has budgeted for other expenses as may be warranted for that organization (graduate student fee remission, relocation expenses, etc.).

Indirect Costs: Indirect rates vary for each organization, and are identified in each of the 16 institutional budgets. JHU charges 63.5% overhead on the first \$25k in value per subcontract; this is a one-time charge and appears in Year 1 of the budget ($15 \text{ subcontracts} \times 63.5\% \times \$25\text{k} = \$238,125$).

Note Regarding NSF Funding Guidelines: Our NVO infrastructure development efforts involve primarily universities but also include some federally funded research institutions. Two of our collaborators, the National Optical Astronomy Observatories and the National Radio Astronomy Observatory, are operated for NSF by university consortia, and one, the Space Telescope Science Institute, is operated by a university consortium for NASA. Participants at the NASA Goddard Space Flight Center are employees of another university consortium, USRA, and of a major IT services corporation, Raytheon. Modest support (3.4% of the total budget) is being requested for work at the Fermi National Accelerator Laboratory (DOE) and the Naval Observatory-Flagstaff Station (DOD); in both cases the funding will be applied toward post-doc salaries.

We understand that NSF funding guidelines strongly favor funding in university-based research environments. In the area of astronomical data management, however, federally funded data centers have been in the leadership role for the past decade, and this is where most of the archived astronomical data already exists, as well as much of the necessary experience and expertise for developing the NVO framework. Despite this fact, over 2/3 of our effort is budgeted for university research groups. In addition, to the largest extent possible, we have planned for work to be carried out by post-docs and graduate students even within the federally funded institutions, in order to maximize the educational benefit of the project. *In no case are funds being requested to underwrite or offset salaries for staff who otherwise have full-time civil service appointments.*

Labor Estimates and Task Allocations: The development efforts for this project will be distributed, just as the NVO itself will ultimately be a highly distributed facility. Our focus is on the NVO *framework*, the suite of interoperability and integration tools and interfaces that are required to support higher level functions and applications. Development and testing of the complete framework requires a large effort, one that goes beyond the funding limits of even a single large-scale proposal. We focus here on the most fundamental framework components. Additional framework elements are the subject of complementary—but independent—proposals (see below).

Our budget is derived from a bottom-up analysis of the various technical work packages, including initial deployment, testing, and evaluation at astronomy data centers. The table below shows, for each area in our work-breakdown structure, the distribution of labor among the collaborating institutions. All labor figures are given in FTE per year and summed over the five-year period of the project.

WBS 1: Project Management efforts are led by JHU/STScI and Caltech.

WBS 2: Systems Architecture work is led by UCSD and Caltech (CACR), groups with extensive experience in the design of large-scale distributed computational systems.

WBS 3: Portals and Workbenches development draws on expertise at several astronomical data centers with sophisticated user interfaces (IRSA, HEASARC) and provides support for portal development for the ground-based astronomy community (NOAO). Integration with workbench applications is led by Caltech (CACR and Astronomy Department).

WBS 4: Metadata Standards work involves many organizations already involved in large-scale data management, both from space missions (STScI, HEASARC, IRSA, Harvard/CXC, NED, ADC) and ground-based surveys (JHU, NRAO, NOAO, UIUC, USNO).

WBS 5: NVO Services development is led by IT centers with similar service layers in grid-based systems (USC, UCSD) with participation from several data providers (IRSA, NED).

WBS 6: Data Models work, like Metadata Standards, involves many data-provider organizations to assure that all aspects of NVO data are accounted for.

WBS 7: Data Access/Resource Layer activities are led by NOAO with support from organizations with similar experience in astronomical data management services or involved in similar efforts for GriPhyN.

WBS 8: Service/Data Provider Implementation and Integration requires modest efforts at all service provider organizations.

WBS 9: Test-Bed developments will leverage on partnerships with the supercomputer centers, coordinated by Caltech (CACR). FNAL and STScI/JHU will also participate, building upon experience with the Sloan Digital Sky Survey and high-end processing systems available at Fermilab.

WBS 10: Science Prototypes will assure that the NVO's science goals are being met by the emerging system infrastructure. JHU, Caltech (Astronomy Department), and U Penn have carried out large-scale surveys and their scientific analysis. ADC and NED bring service integration experience. CMU/U Pitt brings expertise in robust statistical analysis and efficient management of large-scale databases.

WBS 11: Outreach and Education efforts will be led by STScI, which manages the O&E program for the Hubble Space Telescope and NASA's Origins program. The emphasis will be on forming strategic partnerships with external developers of curricula, informal science programs (museums), and general public outreach. Collaboration have already been established with the Association of Science–Technology Centers, Silicon Graphics (Digital Planetarium), Spitz (Electric Sky), the Boston Museum of Science, the International Planetarium Society, the National Air and Space Museum, Project CLEA (Gettysburg College), the Science Education Gateway (UC Berkeley), the Maryland Space Grant Consortium, and the American Museum of Natural History.

	Y1	Y2	Y3	Y4	Y5	Total
<i>1 Project Management</i>						
JHU	0.65	0.65	0.65	0.65	0.65	3.25
STScI	0.90	0.90	0.90	0.90	0.90	4.50
Caltech	1.20	1.20	1.20	1.20	1.20	6.00
	2.75	2.75	2.75	2.75	2.75	13.75
<i>2 Systems Architecture</i>						
SDSC (UCSD)	0.50	0.50	0.20	0.20	0.20	1.60
Caltech	0.30	0.30	0.20	0.20	0.20	1.20
JHU	0.10	0.10	0.10	0.10	0.10	0.50
NOAO	0.10	0.10				0.20
	1.00	1.00	0.50	0.50	0.50	3.50
<i>3 Portals and Workbenches</i>						
Caltech	0.25	0.50	0.75	0.75	0.50	2.75
IRSA/IPAC (Caltech)	0.25	0.50	0.75	0.75	0.50	2.75
HEASARC	0.25	0.50	0.50	0.50	0.50	2.25
NOAO			0.50	0.50	0.50	1.50
	0.75	1.50	2.50	2.50	2.00	9.25
<i>4 Metadata Standards</i>						
STScI (JHU)	0.75	0.50	0.20	0.10	0.10	1.65
CXC (Harvard)	0.35	0.20	0.05			0.60
HEASARC	0.75	0.50	0.20	0.10	0.10	1.65
IRSA/IPAC (Caltech)	0.25	0.20	0.05			0.50
NED (Caltech)	0.25	0.20	0.05			0.50
ADC	0.35	0.25	0.05			0.65
NRAO	0.25	0.20				0.45
NOAO	0.25	0.20	0.05			0.50
USNO	0.15	0.15	0.15			0.45
NCSA (UIUC)	0.90	0.80	0.40	0.30	0.30	2.70
	4.25	3.20	1.20	0.50	0.50	9.65
<i>5 NVO Services</i>						
USC	0.60	1.00	1.00			2.60
JHU	0.20	0.20	0.20			0.60
Caltech	0.10	0.30	0.30	0.10	0.10	0.90
IRSA/IPAC (Caltech)		0.10	0.10	0.10	0.10	0.40
NED (Caltech)		0.10	0.10			0.20
SDSC (UCSD)	0.30	0.50	0.50	0.30	0.30	1.90
	1.20	2.20	2.20	0.50	0.50	6.60
<i>6 Data Models</i>						
CXC (Harvard)	0.50	0.50	0.50	0.25	0.10	1.85
NOAO	0.50	0.50	0.25	0.10	0.10	1.45
IRSA/IPAC (Caltech)	0.50	0.50	0.25	0.10	0.10	1.45
FNAL	0.50	0.50	0.25	0.10	0.00	1.35
U Penn	0.50	0.50	0.25	0.10	0.00	1.35
USNO	0.25	0.25	0.25	0.10	0.00	0.85
NRAO	0.50	0.50	0.25	0.10	0.10	1.45
ADC	0.50	0.50	0.15	0.10	0.10	1.35
	3.75	3.75	2.15	0.95	0.50	11.10

	Y1	Y2	Y3	Y4	Y5	Total
<i>7 Data Access/Resource Layer</i>						
CXC (Harvard)	0.25	0.25	0.25	0.25	0.10	1.10
NOAO	1.00	1.00	0.50	0.25	0.10	2.85
IRSA/IPAC (Caltech)	0.25	0.25	0.25	0.25	0.10	1.10
SDSC (UCSD)	0.50	0.50	0.50	0.25	0.10	1.85
USC	1.00	1.00	0.50	0.25	0.20	2.95
U Wis	0.25	0.25	0.25	0.00	0.00	0.75
	3.25	3.25	2.25	1.25	0.60	10.60
<i>8 Service/Data Provider I&I</i>						
STScI (JHU)	0.20	0.30	0.30	0.30	0.30	1.40
CXC (Harvard)	0.05	0.20	0.30	0.30	0.30	1.15
HEASARC	0.20	0.20	0.30	0.30	0.20	1.20
IRSA/IPAC (Caltech)	0.05	0.20	0.30	0.30	0.20	1.05
NED (Caltech)	0.05	0.20	0.30	0.30	0.20	1.05
ADC	0.15	0.25	0.20	0.20	0.20	1.00
USNO	0.05	0.20	0.20	0.20	0.00	0.65
NRAO	0.05	0.20	0.50	0.50	0.00	1.25
NOAO	0.05	0.20	0.50	0.50	0.20	1.45
USNO	0.15	0.30	0.30	0.30	0.10	1.15
NCSA (UIUC)	0.10	0.20	0.50	0.50	0.20	1.50
	1.10	2.45	3.70	3.70	1.90	12.85
<i>9 Test-Bed</i>						
Caltech	0.25	0.25	0.25	0.25	0.25	1.25
FNAL	0.25	0.25	0.25	0.25	0.25	1.25
JHU				0.20	0.20	0.40
STScI (JHU)	0.25	0.25	0.25	0.25	0.25	1.25
CXC (Harvard)			0.10	0.20	0.25	0.55
HEASARC			0.10	0.20	0.25	0.55
IRSA/IPAC (Caltech)			0.10	0.20	0.25	0.55
NED (Caltech)			0.10	0.20	0.25	0.55
ADC			0.10	0.20	0.20	0.50
NRAO			0.10	0.20	0.25	0.55
NOAO			0.10	0.20	0.25	0.55
NCSA (UIUC)			0.10	0.20	0.25	0.55
U Penn			0.20	0.20	0.25	0.65
	0.75	0.75	1.75	2.75	3.15	9.15
<i>10 Science Prototypes</i>						
JHU	0.50	0.50	0.50	0.50	1.00	3.00
Caltech	0.50	0.50	0.50	0.50	1.00	3.00
U Penn	0.50	0.50	0.50	0.50	0.50	2.50
ADC			0.50	0.50	0.50	1.50
NED (Caltech)			0.50	0.50	0.50	1.50
U Pitt/CMU	0.50	0.50	0.50	0.50	0.50	2.50
	2.00	2.00	3.00	3.00	4.00	14.00
<i>11 Outreach and Education</i>						
STScI	1.00	1.00	1.00	1.00	1.00	5.00
TBD	0.00	0.00	0.00	3.00	4.00	7.00
	1.00	1.00	1.00	4.00	5.00	12.00