

# **Building the Framework for the National Virtual Observatory NSF Cooperative Agreement AST0122449**

## **Management Plan, Revision 2**

**April 2004**

The challenge of building a framework to enable the National Virtual Observatory is being met with a management structure that supports distributed research and development. 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 strong communication and coordination among the distributed, multi-disciplinary, heterogeneous resources, with accountability to both the community and the funding agency. It ensures that astronomy needs drive technology development.

### **1 Management Structure and Key Personnel**

Our team is led by **Principal Investigator and Project Director Dr. Alexander Szalay** (The Johns Hopkins University). The contractual relationship with NSF is via JHU, overseen by Szalay, an astronomer with a strong computer science 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 was a key member of the Sloan Digital Sky Survey project, designing its archive, and a prime mover in the recent National Academy Decadal Survey that strongly recommended an NVO. He ensures that the project is driven by the needs of astronomy. The Principal Investigator/Project Director has overall responsibility for the direction, execution, and completion of the project. He ensures that the project is responsive to NSF and community needs, that it meets its scientific and technical goals, and that its progress and facilities become known to the user community.

**Co-Principal Investigator and Chief Architect Dr. Roy Williams** (California Institute of Technology) is a computational scientist who has been involved in several projects emphasizing high-performance computing, large data and its dissemination and understanding. He is involved with the LIGO software development team and its involvement in the GriPhyn collaboration, the Digital Sky project for astronomical database federation, and the NASA-funded MONTAGE for astronomical image mosaicing. Past work includes the CASA Gigabit Network (1991-95) and the Scalable I/O Initiative (1994-97). Williams is also Applications Lead for the NSF-funded TeraGrid consortium. His main interests are in federating large (TB-scale), heterogeneous databases, and in enabling the interoperation of metadata across semantically diverse systems. As Chief Architect (CA), Williams is responsible for the identification and selection of key technologies needed to support the NVO framework. He also leads in developing partnerships with computational service providers (e.g., national supercomputer centers).

Assisting the PI and CA is a small team of senior scientific and technical managers. The **Project Manager, Dr. Robert Hanisch** (Space Telescope Science Institute), oversees the project schedule and budget, negotiates work packages and task deliverables with all co-investigator organizations, and coordinates the technical development plan with the system architect and other members of the management team. The PM is responsible for ensuring that incremental capabilities are delivered to the end-user/astronomy community. The **Project Scientist, Dr. David De Young** (National Optical Astronomy Observatories) establishes the definition of the scientific objectives of the NVO, based upon a broad community consensus, and is responsible for ensuring that development tasks meet high-level science goals through liaison with the community and by testing science prototypes. **System Architect Dr. Reagan Moore** (San Diego Supercomputer Center) coordinates the overall system engineering and design functions, including interfaces between the architectural elements. The System Architect ensures that the architecture and functionality of the NVO will be structured to realize the scientific objectives of the NVO, and works with the Chief Architect in determining an appropriate set of technologies and interchange standards that will ensure the functionality and interoperability of the different parts of the NVO collaboration.

The education and outreach activities of the project are the responsibility of the **Education and Public Outreach Coordinator, Dr. Carol Christian** (Space Telescope Science Institute). She plans an education and outreach program that capitalizes on emerging NVO capabilities, highlighting both astronomy and information technology EPO opportunities. She establishes partnerships with museums, planetariums, science teachers, scholarly societies, and corporations in order to develop an effective education and outreach program for the NVO initiative.

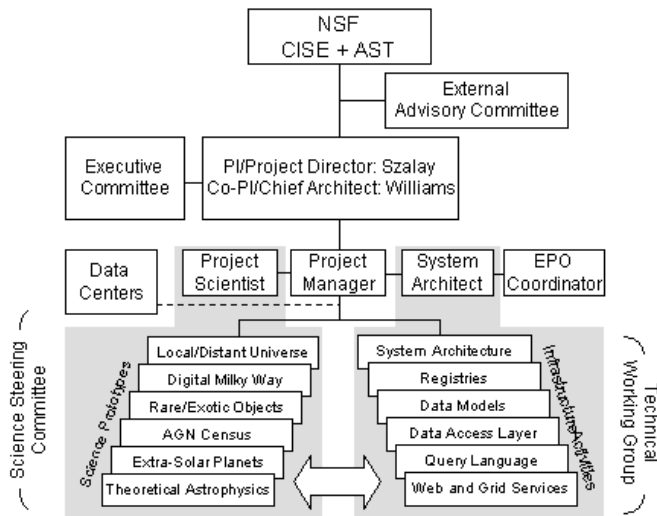
The Principal Investigator chairs an **NVO Executive Committee** comprising the Chief Architect, Project Manager, System Architect, and Project Scientist. The Executive Committee is responsible for approving changes to the project's technical development plan, reviewing progress, and re-allocating funds among tasks as necessary. The Executive Committee supports the PI in overall project leadership, and acts as a body to resolve potential differences among team members. The Executive Committee must discuss and agree upon reallocation of resources among the collaborating organizations. The PI may appoint further representatives of the collaboration or other expert advisors to the Executive Committee as may be required. **Dr. Ethan Schreier** (Space Telescope Science Institute) and **Dr. George Helou** are currently also members of the Executive Committee.

**Development Task Teams** including both IT and astronomy expertise will be responsible for individual work packages. They are organized by technical area and include System Architecture, Registries, Data Models, Data Access Layer, Query Language, and Web and Grid Services. The Executive Committee is responsible for choosing leaders for each Development Task Team, who report to the PM on schedule and budget performance. 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 and the SA. A **Science Steering Committee** including application scientists, the PM, and the SA, reporting to the PS, ensures that the NVO infrastructure meets overall science requirements.

Periodic reviews by an **External Advisory Committee** assure NSF that the NVO framework is responsive to both astronomy and IT community needs. The members of this committee are selected by the Executive Committee in consultation with NSF.

## 2 Project Management

We have successfully met the difficult challenge of assembling a project 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. There are 16 funded organizations (19 when one realizes that there are actually four autonomous groups at Caltech), three unfunded collaborating organizations, external education/outreach collaborators, and international collaborators. Our scientific, technical, and program management expenses account for ~15% of the total project budget, not counting the management efforts of the team and group leads who will have responsibility for design and implementation of specific work packages. 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.



*NVO Management Structure.* The Project Director and Chief Architect, in consultation with an Executive Committee, rely on a Project Manager to oversee the project activities, work packages, and budget. The Project Scientist assures adherence to overall science goals, and System Architect coordinates technical development efforts.

Our project management approach relies upon clearly defined work packages, with aggressive but achievable schedules and agreed-upon deliverables. Participating organizations are held accountable for their contracted work packages, and the Executive Committee has the authority to curtail or terminate participation of groups who do not produce. Individual work packages have terms no longer than one year, with milestones monitored at least quarterly, allowing problem areas to be identified early and resolved without waste of resources.

Communications are extremely important, and frequent group telecons and ~quarterly group meetings are an essential component of the development environment. Collaborative distributed development tools are being utilized as necessary to assure code integrity.

### **3 International Collaboration**

The Virtual Observatory is inherently international. Although the term National Virtual Observatory gained widespread recognition as a top recommendation in the recent Decadal Survey of the United States National Academy of Sciences, there are no borders to astronomy, distributed data systems, and the Internet. There are on-line archives of astronomical data and catalogs in other countries, and international efforts to study federating those archives. We anticipate that, eventually, there will be a Global Virtual Observatory. Our team has established liaison with a number of these international efforts and coordinates our activities with theirs.

In June 2002 the International Virtual Observatory Alliance (IVOA) was formed (<http://www.ivoa.net>). The mission of the IVOA is “to facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems, and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory.” Through international working groups and interest groups the IVOA fosters cooperation among VO development projects and works toward common international standards. The NVO is a founding member of the IVOA and a principal partner among the 14 member projects.

### **4 Overview of Activities by Year**

The high-level goals of the NVO are described in the NVO Project Development Roadmap (Appendix A).

### **5 Work Breakdown Structure**

We have refined our top level WBS structure into sub-activities, and these are described below. Each primary work area has an appointed technical leader who assumes responsibility for the planning, execution, and completion of the tasks in that area. These work area leaders report to the Project Manager and reach mutually agreeable schedules, milestones, and deliverables.

#### **1. Management** (Executive Committee)

A project of this scale, with development efforts distributed over many organizations, requires a substantial investment in project management and coordination.

##### *1.1. General*

General management activities include project planning, reporting, and routine communications, including the organization of team meetings and focused work sessions. We develop and maintain an overall project schedule that reflects available resources and project goals, and identifies project milestones and deliverables. Quarterly and Annual Reports are submitted to NSF describing activities in accord with this Work Breakdown Structure.

##### *1.2. Science*

Science management activities include developing, maintaining, and extending as necessary descriptions of high-level science goals for the Virtual Observatory. These goals are illustrated through specific science scenarios (WBS 2.1).

### *1.3. Technical*

Technical management activities include defining standards for software development and document preparation, and associated configuration management capabilities.

The NVO integrates services and software from many areas of astronomy, and we do not expect large amounts of software to be rewritten according to NVO-specific rules. Rather, we strive for interoperability at the service and interface level, where we care about requests and responses, not implementations. We provide reference implementations for essential services. For efficiency in high-performance applications, it is necessary to link code and libraries, in which case industry-standard practices are being followed.

### *1.4. Financial*

Financial management activities include preparation and oversight of the overall project budget, preparation and monitoring of subawards, review of performance, and reallocation of project resources as deemed necessary by the Executive Committee. Financial reports are provided to NSF on a quarterly and annual basis.

### *1.5. International Coordination and Collaboration*

NVO developments are coordinated with similar initiatives in Europe (AstroGrid, Astrophysical Virtual Observatory), Japan, Australia, India, China, etc., to assure interoperability of the various national initiatives. NVO participates fully in International Virtual Observatory Alliance planning and development. Through the IVOA we collaborate in the development of VO technical standards on the international level. This includes metadata standards and data exchange protocols. We utilize working groups, mail distribution lists, and a collaborative web site (TWiki) to assure maximum interoperability.

## **2. Science Requirements (De Young)**

The development of science prototypes is a staged process that progresses in cadence with the development of capabilities described in other WBS sections and with the increasing awareness of, and demand for, NVO services by the astronomical community. This phased implementation includes the development of demonstration science capabilities at approximately six-month intervals beginning in January 2003. These are not “canned” demonstrations but are applications that provide real functionality and can be used by the astronomical community, with increasing complexity and functionality as time progresses.

NVO science prototypes include capabilities such as a web-based clustering analysis service site aimed at scientific exploitation of massive astronomical datasets, including large catalogs from individual or federated digital sky surveys. The purposes are: (a) demonstrate the kinds of novel science enabled by these massive data sets; (b) serve as a testbed for a variety of clustering algorithms and related methodologies; and (c) provide an immediate, scientifically useful set of tools and services leading to novel scientific discoveries even in this prototype stage. Central to the NVO development activities is support for large-scale cross-correlations between distributed catalogs. We are developing both on-demand and batch-based services for scalable cross-match of astronomical catalogs, utilizing file-based and DBMS-based storage strategies to make the service highly efficient in a number of usage scenarios.

### *2.1. Usage Scenarios*

Develop science use-case scenarios that represent typical end-user requirements and expectations. These scenarios can then be expressed as specific queries.

### *2.2. Requirements Analysis*

Usage scenarios are analyzed to understand their impact on underlying NVO resources and services. Usage scenarios are used to identify the interfaces, protocols, and services needed to implement NVO science.

### *2.3. Demonstration Definition and Review*

Specific science prototypes, or demonstration projects, are being developed and shown to the science community. Prototypes of increasing levels of sophistication and breadth are being developed as the project progresses.

### **3. System Architecture (Moore)**

The NVO system architecture builds upon the technologies coming from the digital library, data grid, persistent archive, and astronomy communities, and upon the grid technology provided in the GriPhyN project. However, the support for federation of multiple data collections, extraction and analysis of data in bulk, and astronomy portals for access will result in a unique NVO architecture. We are using existing data management and collection federation technologies already in development at SDSC. These systems are being used to support the 2MASS collection, and can be extended to support bulk analysis of data from SDSS, DPOSS, and other sky surveys.

The systems architecture supports a bi-directional exchange of computer science technology between the NVO testbed data grid implementation and the existing services of the multiple astronomy sky surveys. The NVO testbed will be used to demonstrate bulk manipulation of images and catalog records across distributed computing and storage resources. The NVO testbed software will include systems for:

- Computational resource management
- Storage management
- Catalog management
- Sky-survey wrappers/mediators for joining existing collections to the testbed
- Information discovery
- Metadata delivery
- Data delivery
- Data manipulation interfaces for
  - o Data model conversion
  - o Analysis services
  - o Correlation services
  - o Presentation services
- Service registration catalog
- NVO portal

The services environment consists of the algorithms for manipulation, display, and correlation. These services are under active development within the NVO partner sites. When a service is ready for installation within the testbed (data model defined, support for data model conversion, demonstration of ability to work at scale, demonstration of ability to interface to data and metadata delivery systems), the service will be made available against all sky surveys that are part of the testbed.

To complete the technology exchange, the NVO data grid will be extended to each NVO partner site through the development of mediators that transform from survey-specific access mechanisms to the NVO data grid access mechanisms. The goal by the end of the project is to support manipulation of NVO data objects directly at each sky-survey site as well as the testbed, and to support execution of NVO services at each of the sky-survey sites as well as the testbed.

#### *3.1. System Design*

The System design specifies the components for managing discovery, access, delivery, and manipulation of data distributed across multiple collections and sites. The system design builds upon the creation of separate infrastructure components for managing data, managing information attributes about the data, and managing relationships between the collection attributes. This makes it possible to separate collection federation development from the underlying data movement and data manipulation development. The system design also creates separate infrastructure for managing the import of collections into the federation, from the management of the federation, and access to the federation.

#### *3.2. Computational Facilities*

Determine nominal computational resource requirements for major data centers participating in the NVO, assuming that each would need to support substantial server-side analysis. Determine computational resource requirements for high-performance computer centers (supercomputer centers) supporting the NVO. Such centers would accommodate the most demanding computation resource requirements. Determine nominal network bandwidth requirements for major data centers participating in the NVO, sufficient to support initial, intermediate, and complex science scenarios. Determine nominal network bandwidth requirements for end-

users of NVO services, sufficient to support initial, intermediate, and complex science scenarios, and note service restrictions that might ensue from limited bandwidth to the end-user.

The security performance requirements strongly depend upon the required transaction rates for interacting with collections. Authentication is required of users to the NVO grid, of users to the collections that are federated by the grid, and between the servers that comprise the grid. Given that NVO connects sites that hold data proprietary for specified time periods after initial collection, the NVO also needs to support authorization and access control lists for individual objects within a collection. This can also include authorization to see attributes used within a collection. The NVO needs to assess the costs of using multiple catalogs to support each type of metadata (discipline, location, authentication, authorization), and whether the performance levels require the integration of these attributes into either one or two catalogs.

### *3.3. Digital Library Integration*

The NVO architecture builds upon standards that are being developed in the digital library community for the management and access of information. The emerging standards are being used to simplify development of grid services, and are being implemented as interfaces to the NVO registered collections. Possible standards include:

- Open Archives Interface – metadata harvesting interface used to support grid registries
- Dublin Core – metadata schema for provenance information used in the grid registry design
- Metadata Encoding and Transmission Standard – metadata schema for the management of descriptive, administrative, structural, and behavioral metadata. METS supports extensions to the schema that can be used to define NVO specific attributes.
- Fedora – data life cycle management technology for the organization of digital entities into a digital library
- DSpace – data life cycle management technology for the organization of digital entities into a digital library

Data grid technology are being integrated with the digital library technology to implement logical name space management and process management for the NVO testbed. Support for four naming conventions is needed within the NVO: global identifier, logical name, physical storage name, and descriptive metadata. The integration of grid process management with the data life cycle management simplifies the development of processing and analysis pipelines.

## **4. Registries (Plante)**

Registries provide the Virtual Observatory with an index, or “yellow pages,” of information resources and computational services that conform to VO query and access protocols. Registries are the foundation upon which VO applications are constructed, enabling those applications to dynamically locate the data and computational services that are needed to carry out the application. Registries can be local to a given organization, containing only local resources or selected distributed resources of particular interest, or can be public and distributed, containing information about many resources and including replicas of complete other registries. Registries collect resource metadata that have been defined in content and structure, and which is represented through XML-based schemas. Registry information is exchanged through agreed-upon publishing, harvesting, and replication protocols. Registry information must also be monitored for quality, integrity, and currency.

### *4.1. Resource Metadata*

In this activity, we define the concepts that describe an NVO resource that are necessary to support automated discovery and utilization of registered resources. These metadata concepts are first defined independent of a particular encoding scheme (e.g. XML, FITS keywords, etc.); as a separate step, we define the mapping to particular encoding schema as needed (WBS 4.2). We first concentrate on the highest level metadata that support the basic registry framework. Later we integrate more specific metadata focused on key types of information (e.g., astronomical coordinates, coverage) and specific services (e.g., data access); this work is done in close collaboration with the Data Models and Data Access Layer WBS components.

### *4.2. Resource Metadata Schema*

This activity focuses on creating the standard XML schemas used to encode the concepts defined as part of WBS 4.1, and is done in collaboration with the IVOA Registry working group. The schemas are tested with the registry prototypes. We are developing related tools to make it easier to use the schemas.

#### *4.3. Publishing and Harvesting Protocols*

A publishing protocol is a mechanism for data providers to publish descriptions of resources they manage into the registry framework so that users can discover and use them. We approach this through what we call “publishing registries” designed specifically for this purpose. A harvesting protocol is a common mechanism for retrieving descriptions from one or more publishing registries into a centralized database. “Searchable registries” harvest the descriptions in order to make them searchable by users and applications. The specifications are being defined in collaboration with the IVOA Registry working group. We are deploying the protocols in our prototype and production registries. As part of this effort, we are also developing tools that make it easy to set up and run the two types of registries.

#### *4.4. Query Protocols*

The query protocol is the mechanism by which applications can submit queries to searchable registries and retrieve descriptions of resources in response. The specification will be defined in collaboration with the IVOA Registry working group. We will deploy the protocol in our prototype and production searchable registries. We are developing an interactive portal for searching the registry as well as tools that help applications do the same on behalf of a user.

#### *4.5. Synchronization, Maintenance, Revision Control, and Curation*

This effort addresses the issues of maintaining a working registry. Synchronization refers to the need to keep registries up to date with the latest available resource descriptions, particularly searchable registries intended to be globally comprehensive. Revision control refers to the problem of how to handle updates to resource descriptions. Curation refers to what an administrator of a registry must do to ensure the integrity and usefulness of the resource descriptions, including how to handle incomplete or incorrect descriptions. We will study these issues in the context of actual operating registries, and we will address them through improvements to the registry protocols and standards, development and publishing of best practices, and administration tools as needed.

### **5. Data Models (McDowell)**

In the NVO context, we must handle heterogeneous datasets from different disciplines and recognize their commonalities. Data models are abstract descriptions of the properties of astronomical data objects (images, spectra, etc.) divorced from both their specific file-based implementations and from the specific metadata used to encode them. The Data Model project will define an extensible and evolving class hierarchy of astronomical data objects covering all branches of astronomy, including data from different wavebands and derived, theoretical, or simulated data. We will define standards for introducing new data models in a flexible way, and set up procedures for the community to communicate such models and agree on standard ones. Our standards will also prescribe ways to handle and propagate unknown data model components, so that specialized analysis clients can make use of extra information recognized only by that (perhaps waveband-specific) analysis system.

We are studying the full range of existing astronomy data models, most of which exist only implicitly in their software systems. We are working with the NVO data centers to generate a catalog of existing data models and their inter-relationships. We will propose, and evolve through community discussion, a generalized abstraction of astronomical data, which we refer to as the NVO Data Model Protocol. This protocol will form the basis for more detailed object models that will support the description of observational, synoptic, catalog, and calibration data and models. It will also support common mechanisms across the different object models to describe relationships and data quality.

The NVO data model is not a software implementation; rather is a “data abstraction protocol” that defines an extensible standard for describing astronomical data sets in a concise and flexible manner, incorporating metadata that allow the data to be self-describing. This protocol provides all of the information that a requestor may need to extract scientifically meaningful information from the data set. The format of data sets that are encoded according to the protocol will be defined so that independent software implementations can be developed that can support this

data abstraction. Part of the NVO data model work is to build a prototype implementation for a subset of astronomical data types in order to validate the data abstraction protocol. In developing the protocol we must ensure that it is sufficiently flexible and well designed to allow other NVO sites to develop their own interfaces without having to use the prototype implementation.

Data models play a key role both in the Data Access Layer (DAL, WBS 6) and in client toolkits for data analysis (WBS 9). Data models govern how metadata is exposed through the VO Registries (WBS 4). All of these components work together to provide a uniform means for dealing with heterogeneous data to permit distributed multi-wavelength data analysis and data mining. Data models and data format mediators also play a key role in the DAL for generating virtual data (e.g., subsetting) to make efficient automated analysis of large datasets possible.

### *5.1. High-Level*

Define general architecture for describing data models. Review existing explicit and implicit data models. Propose generic data model protocol (class definition).

*Image and Spectrum.* These two objects are the core NVO data types, and are critical for the basic NVO demonstrations of data fusion and multi-wavelength analysis. The spectrum model must handle many different wavelength sampling and description methods, as well as cope with (or at least flag) problems such as aperture corrections and non-photometric data. It should also handle single photometric data points as a special case.

*Time Series.* This includes time-resolved photometry, spectrally resolved light curves, and all-sky monitor (synoptic) results.

*Event Lists.* The SAO/CXC and the HEASARC have already done significant work in this area, with successful multi-mission software from both groups relying on common metadata protocols, and this collaboration is being continued. The event list definition will be developed from the CXC event data model and the HEASARC's FITS event list definitions.

*Visibility Data.* Visibility data is used in radio astronomy and other interferometric applications. This is considered a low priority as most NVO users will prefer to access derived image/spectra, as in WBS 6.1, rather than the raw visibilities.

*Catalogs.* Provide an abstract description of the contents of catalogs (including archive metadata) independent of the specific kinds of data they are cataloging. There are two important and different ways in which catalog data are used. First, they can be used as just another form of actual data, since many catalogs contain positions and fluxes that can be incorporated into multi-waveband images and spectra as special cases of the image/spectra objects. Second, they are used as indexes to archival data, and as such contain additional properties describing the methods, protocols, and formats used in querying the attached archives, and descriptions of archive-side computing and data mining capabilities.

*Theoretical Simulations.* The comparison of theoretical simulations and observed data is a complex data integration problem. However at a minimum, for simulations that are specifically targeted to represent a data product generated by a particular instrument, it should be possible to have standard ways to mark the data product as simulated. Describing the parameters of the simulation and handling simulated data that is closer to the universe (e.g., 3D) than to the instrument are much trickier, and should wait until the rest of the project is more mature.

### *5.2. Low-Level*

*Errors and Uncertainties.* Specify data model protocol for describing errors and uncertainties in data. These apply to all data types described in WBS 5.1. Describe nature of errors (Poisson, Gaussian) and methods for error propagation when compatible data sets are combined. Define default criteria for failing to propagate errors; since experience shows that in complicated data manipulations, it may not be possible to provide an automatic algorithm to reliably combine errors, it is important to know when to give up. The protocol will specify what kinds of error and uncertainty information are recognized and understood by the system. This implies constraints on how error and uncertainty information should be presented in the metadata standards.

*Data Quality Measures.* Specify data model protocol for describing data quality of a data object. Quantitative and qualitative data quality measures are present in much astronomical data but there is no overall consistent approach (even when integer quantitative measures are used, sometimes big numbers are good and sometimes bad). We must either map these measures to a consistent approach or provide meta-metadata to describe the

properties of a particular measure. The data model protocol produced will specify what kinds of measure are supported and the ways in which they may be connected to the data objects.

*Relationships.* Develop data model protocol describing relationships among data objects, and defining compound data objects. These may include enriched images that have a data array, error array and bad pixel list, or more complex hierarchies coupling derived products and their progenitors (source list, corresponding collection of images). Support both compounding (treating as single data object) and associating (links between objects).

### *5.3. Descriptors and Ontologies*

Define standard terms (descriptors, known in this context as UCDs or Unified Content Descriptors) for astronomical concepts.

The structural data models describe the roles of metadata in a dataset (array axis, data quality, coordinate system...) and require connection to the astronomical semantics (wavelength, surface brightness, equator...) which will be provided by the standard descriptors. These are used in applications such as cross-matching columns in catalog tables and in expressing queries.

Develop simple methods to compare concepts: wavelength implies frequency, V magnitude is a specific kind of flux, etc. Define the interaction of descriptors and physical units.

Define relationships between concepts (ontologies). Explore the usefulness of software ontologies in the VO context. The initial controlled vocabulary of standard descriptors provides a separation of metadata into classes, but requires human understanding to determine what to do with those classes. A formal ontology would give machine-usable semantics by specifying roles and restrictions of concepts when used within other concepts. Ontology technology is at an early stage of development and is not currently baselined for the initial VO, but to the extent that we can define the standard descriptors formally at this stage, it will be easier to add new technologies later.

### *5.4. Space-Time Coordinates and Regions*

Develop generic data model describing spatial relationships and other data contexts. In the spatial domain, describe survey sky coverage, including support for overlapping regions and excluded regions. Support both celestial and instrumental coordinate systems. Support use of the HTM (Hierarchical Triangular Mesh) as an indexing scheme. Support a superset of the IRAF and CXC region filtering schemes. Provide models and encodings for all astronomical timescales and spectral (including Doppler) coordinate frames to ensure that complete and consistent metadata describing these systems can be represented.

### *5.5. Standard Schema*

Conventions must be established for how data models are to be expressed. Encoding data models into a standard form (i.e., XML) allows applications to parse and use data models using software that is independent of any particular data model terms. This allows for maximal flexibility and extensibility. Rules for the use of namespaces, choice of attributes versus elements, and choices of alternate serializations (VOTABLE versus model-specific schemas) will be provided.

## **6. Data Access Layer (Tody)**

The data access/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 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.

### *6.1. Data Access Services*

At the highest level the NVO data access services provide object-oriented access to the major classifications of astronomical data: source catalogs, 2-dimensional sky images, 3D image cubes, 1-dimensional spectra and SEDs, time series, and so forth. For each such class of data a standard data model and data access protocol is defined. Most data access is mediated on the data server at access time, converting heterogeneous external data into the VO data model implemented by the specific data access protocol used. The same physical dataset may be viewed via different access protocols and data models depending upon the requirements of the analysis being performed. Subsetting and filtering of the data may occur at access time, or synthetic data may be produced. Much data access in the VO is thus referencing *virtual data*. It is this combination of mediation to standard data models plus uniform access protocols that makes large-scale multiwavelength data fusion possible.

In addition, most of the online astronomical data is available through Internet protocols such as FTP and HTTP, and the NVO strengthens support for these by codifying the directories and keywords that are necessary for effective use of them. This codification will take the form of capability documents that are available through the NVO registry.

The NVO registry provides—for each data service or web site—a URI (Universal Resource Identifier) that is modeled on the location of capability documents for the service. As discussed above, this document allows a machine to understand the nature of the request and the menu of response types that is available from the service. It allows, for example, the automatic generation of forms that can be filled in to generate a request.

Data replication and data caching deal with many, but not all, NVO data/processing requests. NVO must be able to deal with arbitrary queries against distributed services efficiently—in real time. Some queries can be accommodated with efficient full-file transfers (Grid FTP) but others will require access to high-performance distributed databases. Development efforts focus on ensuring that Grid FTP meets NVO requirements, and that DB performance is not limited by remote access protocols.

### *6.2. Data Representation*

Each data access protocol defines a query, a query response, and one or more alternative standard data formats in which data can be returned. Each such data format is a representation of the underlying data model for the type of data being accessed (2D image, 1D spectrum, etc.). For example, a 1D spectrum can be returned in XML encoded as a VOTable, as a FITS binary table, or as a CSV text file. Most data access services will also provide a pass-through mechanism to allow access to external data without conversion. The standard data formats will be extensible, allowing data providers to pass along additional collection-specific information and data elements, in addition to those defined by the core model.

### *6.3. Framework (Mediators, Components)*

While the data access protocols are defined in an implementation-independent fashion, as data access gets more complex there is an increased need for reference-grade software that implements the data access services. A data access framework provides reference implementations for both the client and server side of the DAL services, to aid user development of VO-enabled analysis software, and to provide an execution framework for application of server-side analysis and transformation functions.

On the client side this framework will provide a *Virtual Observatory Workbench* that can be used to implement arbitrary analysis applications based on the distributed VO infrastructure. On the server side the framework will provide mediated data access including support for virtual data generation and application of server-side functions and pipelines. Astronomical data processing functionality (mostly from the existing software base outside NVO) will be wrapped as components which are callable by the reference framework, or any other execution environment which supports the standard NVO component interface.

### *6.4. Data Provider/Consumer Implementations*

For NVO to succeed the project must not only provide the NVO infrastructure, it must foster use of this infrastructure by the astronomical community. We are working with data providers to help them implement

data services to publish their data to the NVO, and with the astronomical data analysis community to help them implement applications which use the data access services.

## **7. Query Language (O’Mullane)**

The query language for the Virtual Observatory has been given the name VOQL (Virtual Observatory Query Language). Creation of a new language is a costly and difficult task, thus we have broken this into three layers. The layers also relate directly to ordered phases of development.

### *7.1. Low-Level*

Key information resources in the Virtual Observatory are astronomical catalogs (source lists, observation logs), which are normally stored in tabular formats. Structured Query Language (SQL) provides a ready medium for accessing such information. Many of the catalogs are stored in relational databases, allowing for easy execution of SQL-like statements. The lowest level of VOQL is the Astronomical Data Query Language (ADQL). ADQL is heavily based on SQL with additions to support specification of sky regions and cross-matching of objects between catalogs. This WBS element is concerned with defining the on-wire representation of ADQL and the definition of the interface for accepting VOQL and returning data in VOTable (among other) formats. The catalog publication interface is called OpenSkyNode and will be defined as a SOAP service. The products of this work area include the ADQL and SkyNode specifications as well as prototypes of working SkyNodes in Java and .NET.

### *7.2. Mid-Level*

The OpenSkyNode service supports queries to individual nodes. It also allows us to query the node to understand the structure of its tables. In the next level up we will integrate these nodes into easy-to-use portals. This WBS focuses on a portal that queries the registry to find SkyNodes, and then allows the user to type an SQL-like string (ADQL/string), which is converted to ADQL/XML and sent to one or more nodes. The more ambitious work here is to reformulate SkyQuery.NET to make use of the more generic OpenSkyNodes. This is termed the OpenSkyPortal. Such a portal allows queries to be submitted that include cross matches between catalogs (where a catalog is represented by an OpenSkyNode). We foresee that the ability to perform cross matches may not be supported by all SkyNodes. Hence the OpenSkyNode specification has two categories: Basic and Full, for OpenSkyNodes. The output of this work area will be the more detailed specification of the FullSkyNode within the OpenSkyNode specification., as well as the operational OpenSkyPortal. Another track at this level is the use of the Object Models for querying and returning of data from SkyNodes.

### *7.3. High-Level*

Ultimately we wish to have a high-level, “natural-language” interface for the Virtual Observatory. Here we may express queries such as “Find x-ray observations from multiple catalogs where we have optical counterparts.” This would require interaction with the registry as well as interaction with multiple data sources to achieve. Work in this area is subject to the successful and early completion of the low- and mid-level query language interfaces.

## **8. Web and Grid Services (Williams)**

U. Pitt/CMU, IPAC, and Digital Sky are funded outside the NSF NVO project to develop compute-intensive services with NVO-compliant interfaces. All these organizations work closely with the NVO architecture group to deliver their services to the NVO test bed before public deployment. U. Pitt/CMU will deliver XML-interfaces to their statistical analysis and data-mining tools: the Mixture-Model of Gaussians (FASTMIX) code, Bayes Network anomaly detection software, adaptive kernel density estimators, Cuevos clustering algorithms and non-parametric regression using REACT (funded by NASA AISRP program and the NSF ITR and KDI programs). IPAC/JPL/CACR are developing a custom-access image mosaic engine, which will return an image for a requested size, projection, sampling and coordinate system (NASA HPCC program). The Digital Sky will develop a fast, scaleable cross-comparison and cross-identification version of NED’s Psearch program that will run on SDSC platforms (funded by NPACI).

### *8.1. Web Services*

Web services are the core of the NVO architecture, and we expect that data will increasingly be exposed this way. The NVO uses web services at different levels of sophistication:

- GET/POST services. These have been around for years, and many existing data centers have based their internet exposure on such services. The advantage of using these is familiarity, although they do not have the features of more sophisticated web service protocols.
- SOAP services. These provide self-description through the WSDL file, a true client API, and an exception mechanism. SOAP services are designed for computer-computer interaction, whereas GET/POST services are built for a human filling in a form that is sent to a computer.

The mission of the NVO is to define the nature of these services both semantically and formally, so that they form a set of standard components to which portals and client software can connect.

### *8.2. Grid Services*

The GriPhyN project is developing technology to support the execution of acyclic directed graphs of processes. For a given task, the underlying processes are identified, along with the required data files. An execution plan is developed that is infrastructure independent. A job execution manager maps the execution plan onto available compute resources, and tracks the completion of the multiple processes.

The NVO grid will build upon the process management technology that is being developed within the GriPhyN project. The technology will first be implemented as part of the NSF TeraGrid, and placed into production. Once the multiple NSF TeraGrid sites (SDSC, Caltech, NCSA, and ANL) demonstrate robust use of the technology, we will employ the same capabilities within the NVO grid. The initial support mechanisms for computation within the NVO grid will be based upon the current job-oriented management environment supported within the NSF PACI centers. Management for services provided over the web will be restricted to the control policies implemented at each web service site, with no coordination made between the individual services.

The NVO grid will build upon the authentication technology that is being developed within the NSF Middleware Initiative, which is based upon the Grid Security Infrastructure. The technology will first be implemented as part of the NSF TeraGrid, and placed into production. Once the multiple NSF TeraGrid sites (SDSC, Caltech, NCSA, and ANL) demonstrate robust use of the technology, we will employ the same capabilities within the NVO grid. Alternate technology for authentication is provided by the SRB (challenge-response mechanism similar to Web technology). The NVO grid will conduct assessments of the performance of the authentication system. The major requirement is that the authentication mechanism must meet the transaction rate needed for bulk processing of data. If aggregation is used to minimize the transaction rate, the authentication mechanism will need to support authentication to the aggregated data without requiring a separate authentication step for each digital object in the aggregated data.

The NVO grid will build upon the storage environment that is being implemented within the NSF TeraGrid. The TeraGrid will provide over 10 PB of storage, with over 500 TB of disk cache. The management mechanisms for storage will be tied to resource allocations for NSF peer reviewed research. Since the size of the current NVO digital holdings is on the order of 50 TB, the NVO project will establish a small allocation with the TeraGrid for use of storage resources. Once the multiple NSF TeraGrid sites (SDSC, Caltech, NCSA, and ANL) demonstrate robust use of the storage systems, we will employ the same capabilities within the NVO grid. The NVO grid will conduct assessments of the performance of the data caches. The major requirement is that the data access rate must meet the streaming rate needed for bulk processing of data. We will initially keep at least two 10-TB collections on-line on TeraGrid disk caches. We will also propose the archiving of the NVO testbed collections as replicas within TeraGrid archives, to enable migration of collections onto disk cache for bulk analysis. This will off-load bulk processing from other NVO sites, if so requested.

### *8.3. Computational Resource Management*

We will deploy an end-to-end, operational system that gives NVO users access to compute-intensive grid resources as an integral part of archive access and data fusion portals. This system includes a set of Grid-based “agent” routines that provide the control infrastructure for each of the low-level supercomputing applications. Agents initialize requests and report status. We will develop the interfaces between these agents and grid tools such as Condor and Globus. Eventually, we envisage a richer set of functions, such as providing the means for restarting a request that has stopped or has been partially filled.

#### *8.4. Virtual Data*

The NSF-funded GriPhyN project (Grid Physics Network) is researching the software infrastructure that can support “virtual data”—dynamic creation of large datasets with efficient replica management. One of the physics projects funded is the Sloan Digital Sky Survey, a close collaborator with the NVO. We will leverage this technology throughout the NVO, and exemplify it in the testbed. The GriPhyN project is also investigating alternate approaches for supporting massive metadata requests against collections (the standard 20 queries created for SDSS). The queries constitute services that can be run for the user, for which it is possible to store results.

Virtual data is data created on demand, or on-the-fly, in response to a query that requires server-side processing. Virtual data primarily consists of the set of attributes needed to manage execution of the original algorithm (parameters and input file), and an interface to drive the processing steps. Data subsetting is a simple version of this capability. Another simple version is storing mosaic result sets.

Complex science queries require server-side storage of intermediate query results and intermediate data products (image subsets, image mosaics, spectral energy distributions). NVO data centers and computational service centers must be able to accommodate temporary storage of these intermediate results, treating them as real entities so far as the user is concerned. The NVO must also be able to dispose of these intermediate products once they are no longer needed.

#### *8.5. Application and Service Integration with the Grid*

The NVO is working closely with the NSF TeraGrid project and other projects that are directed at building Grid infrastructure. Several NVO-sponsored grid projects are being deployed on the TeraGrid, and there is now a registered collection of data services that these applications will draw on. The NVO will also specify standard interfaces to persistent, authenticated services that do computing and data mining. The NVO and other projects will connect such services into a cyberinfrastructure, i.e., a workflow of services that can be published and discovered in a registry.

### **9. Applications (McGlynn)**

Efforts in this WBS element are directed towards providing the actual web sites and tools through which users will interact with the Virtual Observatory. The goal of the NVO is to enable easy access to astrophysics information to all users. To the extent that this goal is eventually realized there will be many NVO applications, some of which are generic access portals and others are specialized. Within the project we are building several broad portals that enable access to major constituents of the NVO. Much of the work in this area, however, is in establishing communications and interfaces with groups developing NVO-compliant applications outside of the core project.

#### *9.1. Data Location Services*

Previous data location and integration services such as SkyView and Astrobrowse provided simple discovery tools for time and position queryable astronomical resources. A new Data Inventory Service is being developed that uses the full metadata resources of the NVO to allow users to discover and access resources using the resource metadata descriptions available in the Registry. This Data Inventory Service is then instantaneously extensible to any new VO-compatible resources.

#### *9.2. Cross-Correlation Facilities*

We are investigating scientific applications of massive cross-match using both on-demand and batch-based services for scalable cross-match of astronomical catalogs. We are developing file-based and DBMS-based storage strategies to make the service highly efficient in a number of usage scenarios.

#### *9.3. Image Combination and Registration*

We are utilizing services that facilitate the combination of catalog and image data from multiple sources. Mosaicing tools add datasets with well-defined world coordinate systems. Registration tools define and refine the world coordinate systems of datasets where the WCS is not provided explicitly, or may have substantial uncertainties. Such tools may use the well-defined locations of catalog objects within the field of view to establish the coordinate frame. Tools for the direct cross-registration of datasets are also of interest so that even

if the underlying coordinate system is not known precisely, it is still possible to compare the data from multiple sources.

#### *9.4. Visualization Tools and Services*

We will create a clustering analysis service for multi-parameter datasets, aimed at scientific exploitation of massive astronomical datasets, including large catalogs from individual or federated digital sky surveys. This service will allow interactive, high-performance visualization of datasets in large-dimension abstract spaces. The purposes are: (1) a demonstration of the kinds of novel science enabled by these massive data sets, (2) to serve as a testbed for a variety of clustering algorithms and related methodologies, and (3) to provide an immediate, scientifically useful set of tools and services leading to novel scientific discoveries even in this prototype stage.

#### *9.5. Theoretical Models*

In consultation with the theoretical astrophysics community, we are developing a set of NVO capabilities that will be of particular use to the theory community. In addition, the theory community will develop a series of science prototypes and large scale datasets resulting from numerical simulations that will be of general interest to astronomers for comparison with large-scale observational datasets. The tools and capabilities needed to make such comparisons are being developed in parallel with the dataset definition and creation, and the capabilities made available to the community will begin with modest examples in Year 2 and progress to larger and more complex datasets and tools in the following years. In addition, because theoretical simulations produce many spectacular and illustrative graphics and animations, special effort will be made to select useful subsets of data that will be ideal for use in public outreach activities. These goals will be accomplished via the establishment of technical working groups within the theory community and through workshops to develop and refine the efforts as the need arises.

#### *9.6. Statistical Analysis*

Tools that provide standard statistical tests and analyses of datasets will be extremely valuable for building analysis pipelines within the VO. Simple tools in this area include statistics like the calculations of the averages, medians, and the like. Interfaces to linear-least squares and more sophisticated fitting algorithms, chi-squared and KS significance tests, cross- and auto-correlation tools will also be needed. In many cases these tools may be providing wrappers to existing statistics packages in the community, though some may be simple enough that recording them will be desirable. This is especially true when by performing the simple tasks near the source of the data, the data transfer requirements for the pipeline as a whole can be substantially reduced.

#### *9.7. Data Mining and Outlier Identification*

Data mining tools are essential to enabling scientists to work with the very large datasets within the Virtual Observatory. Data mining tools enable users to sample data according to complex criteria and to look for patterns and unusual elements within the data. Data mining tools include elements of database queries, cluster analysis, classification and visualization. This area is undergoing substantial growth in both the commercial and scientific arenas. As part of this project we anticipate providing standard tools for identifying outliers in large data sets and to encourage the incorporation of more sophisticated data mining algorithms in VO compatible wrappers.

#### *9.8. Interfaces To and From Legacy Software Systems*

In order to support distributed data analysis and data mining applications, interfaces will need to be developed to existing data analysis and data mining tools and packages. For example, NOAO would interface IRAF to NVO using the data access layer (WBS 6). Similar interfaces will be developed for radio astronomy analysis software, x-ray data analysis software, and statistical/data mining tools. This will be useful both for the multi-wavelength data analysis capability provided, and as a testbed for end-to-end testing of the data access portal and general distributed data analysis using NVO. When the data access layer and data models are sufficiently developed we will install the software at our principal development sites and interface it to our local archives and selected analysis packages. We will then be able to conduct interoperability tests and demonstrate a true distributed multi-wavelength data analysis capability. In the later phases of the project we will extend the testbed to the supercomputer centers and the TeraGrid to demonstrate scalability for large scale distributed multi-wavelength data mining.

## **10. Community Engagement** (Hanisch)

### *10.1. Documentation*

The NVO Project maintains a managed document repository where documents of various sorts—status reports, technical specifications, presentations, meeting minutes, etc.—are stored. Upload privileges are given to team members, and the collection is open to anyone visiting the NVO web site. Multiple document formats are accepted, and to the extent possible standard portable document formats such as PDF are provided.

### *10.2. Web Site*

The NVO project maintains a web site for 1) facilitating communications among team members, and 2) informing the community about technical and scientific progress. The web site provides links to documentation, interfaces, software, and applications.

### *10.3. Technical Training Initiatives*

The astronomy community needs to be trained in using the new NVO tools and interfaces so that applications and services can be developed outside of the core project. Once a core set of services and interfaces have been implemented, we will provide seminars and training sessions in a variety of forums such as ADASS Conferences, AAS Meetings, and dedicated Summer Schools.

### *10.4. Advocacy*

We maintain communications with US funding agencies, continuing to clarify NVO goals and define the NVO mission. Primary contacts are NSF and NASA. In addition, we work to inform the community about NVO goals, accomplishments, and plans through our participation in national and international conferences (AAS, SPIE, IAU, etc.).

## **11. Education and Public Outreach** (Christian)

Many education and outreach organizations have expressed interest in NVO. We intend to capitalize on that enthusiasm by developing partnerships with organizations that have extensive experience with development of education and outreach products and existing dissemination pathways that will ensure widespread distribution of NVO outreach materials. We are forming partnerships with the outreach organizations of other NASA missions and ground-based astronomy facilities, with professional societies (AAS, ASP, AAPT, etc.), with developers of astronomy-related commercial software (e.g., Starry Night), with science museums (e.g., AMNH/Hayden Planetarium), and with textbook publishers and curriculum developers for K-12 and undergraduate education. Our partners tell us what capabilities they desire from NVO in order to have an effective and widespread impact on the public, and we use these recommendations to define a set of outreach requirements for NVO infrastructure development. Once the recommendations have been incorporated into NVO, we will host workshops and provide interface documentation for all educators who wish to develop education and outreach products based on NVO. In the meantime, we will continue to seek funding for in-house development of educational and outreach resources.

### *11.1. Strategic Partnerships*

*NASA Missions.* Engage education and outreach arms of NASA missions in development of products related to NVO. Collect feedback on how outreach interface can meet their needs. Coordinate outreach activities between institutions so as to minimize duplication of effort.

*NSF Facilities.* Engage education and outreach arms of ground-based facilities in development of products related to NVO. Collect feedback on how outreach interface can meet their needs. Coordinate outreach activities between institutions so as to minimize duplication of effort.

*Professional Societies.* Recruit potential outreach partners via professional societies and disseminate information about NVO outreach at meetings of those societies. Organize sessions at AAS, ASP, and other relevant meetings on doing outreach with NVO.

*Commercial Software Developers.* Encourage developers of astronomy-related commercial software products to draw on NVO content. Solicit feedback on their interface needs. Provide interface documents for software developers explaining how to link commercial products with the NVO interface.

*Science Museums.* Develop partnerships with museums and planetaria interesting in developing interactive exhibits based on NVO. Create interactive NVO kiosk in partnership with one or more of these institutions.

*Textbook Publishers and Curriculum Developers.* Stimulate interest in NVO among curriculum developers. Work with interested parties to develop classroom activities based on NVO.

#### *11.2. Formal Education*

Seek funding to develop K-12 curriculum products based on NVO. Adapt existing education products to incorporate NVO activities.

#### *11.3. Informal Education*

Develop NVO kiosk with museum partners. Seek funding to develop additional museum-related products based on NVO. Adapt existing informal-science products to incorporate NVO.

#### *11.4. Outreach and Press Activities*

*Public NVO Portal.* Many public NVO portals are likely to appear, with or without our encouragement. We will develop a portal providing access to all the NVO-related materials developed by our partners.

*Amateur Astronomers.* Create and publicize opportunities for amateur astronomers to submit their own observations to NVO. The best of these should be flagged to be of significant interest to the public.

*Press Releases.* STScI's Office of Public Outreach and JHU will produce press releases on behalf of NVO. Our goal is to produce at least one NVO release per year.

#### *11.5. Technical Development*

Access to education and outreach products will require special interfaces that either pre-process or select from prepared materials those appropriate for E&O use. We need to avoid inadvertent delivery of, e.g., a 4k x 4k x 4k FITS format data cube, to an elementary school teacher looking for GIF pictures or press releases.