

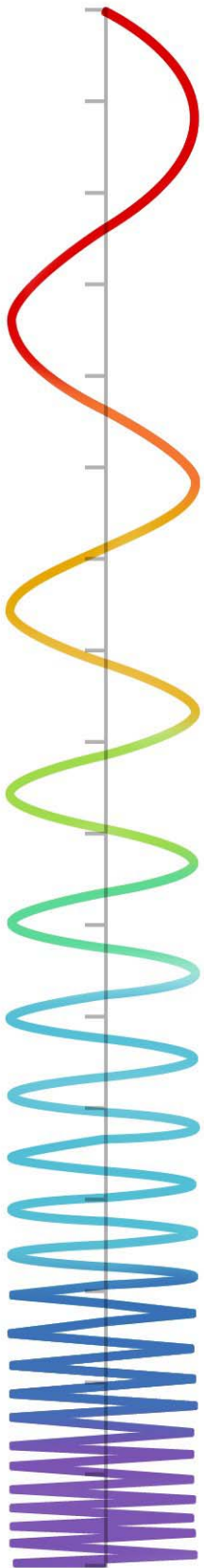
Annual Report  
October 2002—September 2003

Building the Framework for the  
National Virtual Observatory

NSF Cooperative Agreement  
AST0122449



INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE



Executive Summary .....	1
Activities by WBS .....	3
1 Management.....	3
2 Data Models .....	4
3 Metadata Standards.....	5
4 Systems Architecture .....	9
5 Data Access/Resource Layer .....	14
6 NVO Services .....	18
7 Service/Data Provider Implementation and Integration .....	20
8 Portals and Workbenches.....	21
9 Test-Bed.....	22
10 Science Prototypes.....	22
11 Outreach and Education.....	27
Activities by Organization .....	28
Caltech–Astronomy Department .....	28
Caltech–Center for Advanced Computational Research (CACR).....	28
Caltech–Infrared Processing and Analysis Center (IPAC).....	30
Canadian Astronomy Data Centre/Canadian Virtual Observatory.....	32
Carnegie-Mellon University/University of Pittsburgh (CMU/UPitt) .....	33
Fermi National Accelerator Laboratory (FNAL).....	33
High Energy Astrophysics Science Archive Research Center (HEASARC).....	34
Johns Hopkins University.....	35
Microsoft Research .....	35
National Optical Astronomy Observatories (NOAO).....	35
National Radio Astronomy Observatory (NRAO) .....	36
Raytheon/ADC.....	36
San Diego Supercomputer Center.....	37
Smithsonian Astrophysical Observatory.....	38
Space Telescope Science Institute .....	38
United States Naval Observatory.....	40
University of Illinois-Urbana/Champaign/National Center for Supercomputer Applications (UIUC/NCSA).....	40
University of Pennsylvania .....	41
University of Southern California (USC/ISI) .....	42
University of Wisconsin .....	43
Participant Report .....	44
Publications.....	59
Virtual Observatory Articles in the Popular and Technical Press .....	62
Acronymns.....	63
NVO Project Roadmap .....	66

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

**Period covered by this report:** 1 October 2002 - 30 September 2003  
**Submitted by:** Dr. Robert Hanisch (STScI), Project Manager

**Executive Summary**

The early technical developments in the first year of this project led to a successful round of science demonstration project in Year 2, culminating in the NVO's first general science application, the Data Inventory Service. The NVO project continues to play a major role in fostering international collaboration through the International Virtual Observatory Alliance. Technical and scientific progress to date has helped us to recraft our project roadmap, a primary focus of which is to bring well-documented and tested VO development tools to the astronomical community by the summer of 2004. Similarly, a small number of education and outreach pilot projects will help us to validate our EPO approach.

*NVO Science.* The three science demonstration projects selected in Year 1 led to successful prototypes shown at the January 2003 AAS meeting in Seattle. These included:

- A brown dwarf candidate search, based on a cross-correlation between the 2MASS and SDSS Early Data Release source catalogs. The demonstration not only confirmed the known brown dwarfs in the region of overlap of the surveys (in a matter of minutes rather than months) but also located several new candidates, one of which was confirmed spectroscopically.
- A galaxy morphology analysis program, in which the morphological parameters of galaxies in rich clusters were measured dynamically with an algorithm deployed to a computational grid.
- A gamma-ray burst follow-up service, where images, pointed observations, and cataloged measurements of an arbitrary location on the sky were assembled on demand using standard NVO protocols. This multiwavelength data collection tool was later refined into a generic Data Inventory Service and released for general use following the IAU General Assembly in Sydney (July 2003).

Planning began for our January 2004 science demonstrations, focusing for the first time on the integration of theoretical simulation data with the VO.

*NVO Technology.* Following on the initial success of the VOTable format specification and development of associated I/O libraries, we enriched the suite of NVO data access tools with fully functional Cone Search services (catalog access) and the Simple Image Access Protocol. Cone Search and SIAP services were utilized to implement the science demonstration projects, and by the end of project Year 2 ~100 Cone Search and SIAP

services were listed in our prototype registry. The registry collects metadata about NVO resources (data collections, observation logs, catalogs, computational services) using internationally agreed upon metadata definitions, encoded in XML schema. We experimented with Web Services, for example, in building wrappers for prototype registry queries, and gained considerable experience in both building and consuming from Web Services. NVO team members both led and participated in the several IVOA technical working groups: Registries, Data Models, Data Access Layer, Unified Content Descriptors, VO Query Language, Grid and Web Services, and VOTable. In addition, NVO personnel co-lead the development of an IVOA standards process, a process recently endorsed by the IVOA Executive Committee.

*NVO Project.* The NVO Project is operating in full gear. More than 70 people at 20 organizations and institutions are participating in the project, with a full-time equivalent level of effort of approximately 17 persons. The project is carried out with less than 10% of the resources allocated to management activities.

Funding underruns from Year 1 (owing to lags in start-up and hiring) continue to provide a financial cushion and will allow for work to continue at a strong level throughout project Year 3 and into Year 4.

Community interest in NVO continues to grow. Scientists shown the NVO demonstrations at AAS meetings react with enthusiasm, and observatory and project leaders have become more VO-aware, with a greater emphasis on VO-ready archives of processed and documented data. The astronomical software development community has expressed incredible interest in VO tools and technologies, with over 200 people attending a VO Software Tutorial organized in conjunction with this year's ADASS Conference. The latest international VO "interoperability" workshop attracted 120 participants.

## Activities by WBS

### 1 Management

#### 1.1 Science Oversight

The Executive Committee takes an active role in oversight of the team's science demonstration projects. The Project Scientist, D. De Young, plays a particularly important role in this regard, keeping the demonstrations focused on research capabilities. Our three initial science demonstrations (brown dwarf search, galaxy morphology analysis, and gamma-ray burst follow-up) showed the variety of applications that will be supported through the NVO. De Young, with the assistance of P. Teuben (U. Maryland), has been defining the goals for a new science demonstration in which we will bring together theoretical simulations (of globular clusters) with observational data.

D. De Young, G. Fabbiano, and R. Hanisch continue as members of the Astrophysical Virtual Observatory Science Working Group. We hope to capitalize on some of the AVO work to provide convenient graphical browsing capabilities for deep surveys.

A. Szalay has led the development and distributed installation of SkyNodes, which are a standardized interface for publishing database queries to the network. SkyNodes can participate in a network of distributed databases, supporting complex cross-correlation requests in an optimal manner. The SkyNode/SkyQuery design is being generalized and is expected to be adopted throughout the VO community.

#### 1.2 Technical Oversight

The Executive Committee also is actively involved in technical discussions, e.g., through the regular telecon meetings of the Metadata Working Group, the biweekly Project Status Review telecons, and leadership and participation in the various IVOA working groups. EC members have been influential in the development of the Resource Metadata standards, Unified Content Descriptors, and VO Query Language.

This past summer we took a close look at the relationship between NVO technology and the Grid. Most all VO technologies have Grid counterparts, though in many cases the Grid components have not stabilized or are in quite early stages of development. The EC continues to monitor Grid developments closely, and will incorporate relevant Grid technologies into the VO framework as they mature and stabilize.

#### 1.3 Project and Budget Oversight

The NVO EC undertook a revision of the project Roadmap in the past several months (the revised Roadmap appears as an appendix to this report). The Roadmap identifies key goals in the areas of science, technology, and education/outreach, and thus provides both a general sense of direction for the project and a yardstick for measuring success.

We have found that the work breakdown structure that we constructed at the beginning of the project is overly complex and at least partly redundant. This has complicated project reporting and confused WBS leaders, who were unable to clearly distinguish tasks in their areas of responsibility. It has also led to a project schedule that was too detailed to use in a practical way. In the beginning of project Year 3 we intend to revise the WBS and update the project schedule, and will submit these revisions to NSF for review.

The project Education and Outreach Coordinator, Dr. Mark Voit, left STScI in July 2003. It was not practical for him to continue in this role in his new position at Michigan State University. Dr. Frank Summers, also of STScI's Office of Public Outreach, assumed the role of EPO Coordinator. Frank brings to the project his extensive experience in informal science education and science visualization.

The project is in good shape financially. First-year cost underruns (owing to delays in start-up and hiring) have been carried forward to allow us to sustain full staffing throughout Year 2 and Year 3. We retain a modest contingency fund and have been able to deal with some work reallocations among team member organizations effectively.

## **2 Data Models**

### *2.1 Data Models / Data Model Architecture*

In the past year we achieved consensus on an international adoption process for data models, pending agreement on a general IVOA standardization process. DM work packages were established under an IVOA DM Working Group. Progress toward a data model definition with consensus from the IVOA group has been made.

A concept paper describing parameterized content descriptors and presenting an initial organization of the problem domain was circulated by J. McDowell (SAO) on the IVOA mailing lists. Data modeling discussions are archived on the [ivoa.net](mailto:dm@ivoa.net) mailing list [dm@ivoa.net](mailto:dm@ivoa.net).

J. McDowell led the IVOA DM Working Group meeting in Cambridge, May 2003. Progress is now being made on the small scale (Quantity model) and large scale (Observation model).

Data Model group members also participated in the UCD redefinition effort. The UCD2 proposal provides an intermediate step between the original UCDs and a precise data model, giving a structured and comprehensive description of the astronomy problem domain but with less precision than needed for data analysis.

A discussion on ontologies for scientific units was carried out on the data model mailing list.

### *2.2 Data Models / Data Types*

Discussions continued regarding the data models for spectral datasets. We analyzed the implications of the spectral use case survey on spectral data models. The need to describe different kinds of observables was emphasized, and some special cases highlighted. A document was produced to summarize the proposed alternatives for the Quantity data model.

### *2.3 Data Models / Data Associations*

Work on generalized coverage (bandpass, regions etc) is ongoing; the Metadata Working Group (see WBS 3.5) is taking the lead on regions specifications.

S. Lowe (SAO) is working on a description of image mappings and coordinate systems.

## **3 Metadata Standards**

### *3.1 Metadata Standards / Basic Profile Elements*

The first version of the Space-Time Coordinates (STC) metadata definition has been published. This includes spectral bandpass coverage and spatial region specifications. This version is being integrated into the data model and registry descriptions. It is clear that incremental improvements and adjustments will be required as we gather more experience with usage. We are currently considering deriving a restricted subset or an implementation level categorization for use in registries. The design of the system of coordinate projection metadata is to be taken up with AstroGrid/Starlink colleagues.

There are some concerns that the STC may be too complicated (in terms of allowable options) for use in resource registry entries. The mapping of the coordinate systems to the data has not yet made progress.

### *3.2 Specific Profile Implementations*

The development of implementation for the generic resource profile was a major topic of discussion in the Metadata Working Group. The relationship of the generic resource profile to the profiles for specific kinds of service, e.g., Cone Search and SIAP services, was debated with and a hierarchical description of resources was continually refined.

The current UCD framework was examined for appropriateness with regard to the kinds of observational tables used in active mission archives. The current framework lacked direct support for quantities associated with the proposal and reduction process. A report discussing these issues was circulated.

During the Victoria team meeting and thereafter there was extensive discussion of the relationship of data set identifiers proposed for use by NASA data centers with the VO identifier framework. After extensive discussion the consensus was that with some care

these identifiers could and should be accommodated within the large VO framework. These are proposed for use beginning in early 2004.

At the HEASARC efforts were begun to specify the resource metadata (using the then current RM document) for all HEASARC tables, approximately 300. The resulting information is now undergoing internal review by HEASARC scientists.

The relationship of the ADEC dataset identifiers with the VO identifier framework should continue to be monitored. The continued evolution of the basic resource metadata means that metadata resources already in place may need at least modest revision.

### *3.3 Metadata Representations and Encoding*

*Schema Definition Framework.* This year, we have closely examined the question of how we define and publish metadata standards, using resource metadata (see WBS 3.4) as our testbed. Our first year demonstrations (completed and presented in January 2003) were instrumental in understanding not only what kinds of information are needed, but also the need for consistent representation of that information. Consistent with the process outlined by the Data Models group, metadata definition starts with a prose document defining the essential concepts, giving each a name. This was realized in the form of the “Resource Metadata for the VO” document (Hanisch et al.; see WBS 3.4). From that document, we derived an XML Schema, known as VOResource, for encoding this metadata in XML. Through the XML modeling process and subsequent use in the registry prototypes, we discovered where additional concepts and other improvements were needed, which fed back into later revisions of the “Resource Metadata” document.

From the resource metadata prototyping, we developed a general approach to encoding metadata in XML with the goal of achieving good clarity, straightforward reusability and extensibility, and ease of use with widely available XML tools. We outlined a object-oriented approach for representing extensible concepts like “Resource” which we can extend to describe more specific types of resources such as “services”, “data collections”, “organizations”, etc. We demonstrated how we can use XSL to create metadata dictionaries and convert into other forms, such as Dublin Core (<http://www.dublincore.org>), and we gained considerable experience using various code-generating toolkits in real applications.

Global interoperability of metadata is essential, and thus much of the development we have done within the NVO project has been passed to the global IVOA forum for further revision to address broader needs. This has inevitably illustrated differences in approach between the various VO projects. One issue we have had to address with regard to the resource metadata is whether metadata definitions should be fuzzy, allowing them to span across diverse resources, or precise, to support unambiguous interpretation needed by processing. At the moment, there is not a universal answer to this issue, and so we address it on a case-by-case basis; however, we know that the lower the level of the metadata, the more precise it needs to be.

We still debate the issue of what can be registered in our registries. While we are concentrating on a small set of resource types for the short term (data collections, services, and organizations), there is a question of what new kinds of resources we allow in the future (e.g. people, standards). This affects the foundation of the resource metadata model.

Finally, we are still working to understand how metadata consumers can cope with new extensions of the standard metadata. The OAI concept of metadata formats is perhaps the best scheme for exposing support for extensions.

*Naming Standards, DOIs.* In collaboration with the NVO Metadata Working Group and the IVOA Registry Working Group, R. Plante led the development of the IVOA Identifiers Working Draft (<http://www.ivoa.net/Documents/>). These identifiers are used to uniquely name resources within a global registry. Based on the IETF standard for URIs (Berners-Lee et al. 1998, RFC 2396), the work draft defines an identifier in terms of two components: an authority identifier that establishes a namespace, and a resource key that identifies the resource within that namespace. It also defines two encoding forms: an XML format and a URI format.

We also coordinated with a joint effort between the astronomical journals and the NASA Astrophysics Data Centers Executive Committee (ADEC) to establish persistent links between published articles and the datasets they are based on. This coordination led to modifications to the identifier specification that allowed ADEC to adopt the IVOA identifier syntax.

The IVOA identifiers are nominally organization-dependent and can distinguish between separately curated versions of the same data collection. However, the ADEC effort to support data links from the literature drives the need for persistent *logical identifiers*. Support for this in the VO needs to be based both on the IVOA Identifiers standard as well on the standards for Registries, which are still in development. A proposal for encoding logical identifiers in the Resource metadata that interoperates with the ADEC effort has been proposed.

### 3.4 Profile Applications

*Query Profiles.* Early work on XML-based query languages was begun this year by E. Shaya and B. Thomas. They envisioned two levels of languages: a high-level, science-oriented language used by end-user clients and a low-level, data-oriented language used to interrogate specific datasets from an archive. When this work was transferred to the IVOA VO Query Language (VOQL) working group, it was integrated with the SkyNode effort at Johns Hopkins. A three-level approach was adopted where the lowest level, now referred to as the Astronomical Dataset Query Language (ADQL), is being developed first. ADQL is being standardized within the context of an OpenSkyNode Working Draft. It is based on an-SQL data model but is expressed in XML. In particular, the query constraints are represented as a fully tagged, parsed tree; this greatly eases the conversion to local query forms required internally by a data provider's database.

There is a question regarding the role of UCDs in ADQL and the SkyNode interface. From one perspective, the client interacts with a remote table via the local column names; the client learns what concepts these names correspond to prior to an actual query by pre-analyzing the UCDs returned in a special metadata query. From another perspective, the client can refer directly to columns in a query by their UCD names; in this case, the server takes responsibility for mapping the UCD names into local column names. In either case, it will be important to expose the local names and descriptions to the client as these will be more specific than the fuzzy UCD association.

*Service Directory Profile.* A major focus this year has been on defining the metadata for describing resources. This metadata allows data and service providers to publish descriptions of their resources into registries; users can search these registries to discover relevant resources. The metadata definition takes two forms. The first is the “Resource Metadata for the VO” Working Draft (WD-RM; <http://www.ivoa.net/Documents/>), which defines the metadata independent of its encoding. It focuses on concepts common to all resources: what the resource is, its type, who curates it, and its target audience. In particular, we have integrated concepts for identifying resources useful for education and public outreach. It also looks at concepts specific to services.

Second is the derivative XML Schema form, VOResource (see <http://www.ivoa.net/wiki/bin/view/IVOA/IVAORegWp03>). It incorporates the WD-RM metadata into an object model: specific resources classes, such as Organization, DataCollection, and Service, subclass from the generic Resource, thereby inheriting the Resource metadata; within these subclasses, more specific metadata are added. We have applied this model to describe Cone Search and Simple Image Access services, including their inputs and the columns found in their output VOTables. This schema was used in the NVO Registry prototype used to support the Data Inventory Service.

A question remains as to what kinds of things we should allow to be registered—in other words, what we define to be a resource. The most striking example is whether we should consider a “Person” as a resource we might look-up in a registry (e.g. to discover a person’s access rights); however, this issue applies more subtly to other entities. The essential issue here are whether these potential resource types are “curatable”—that is, we can identify someone responsible for the resource.

Another important issue we are still studying is one of granularity in our registries: that is, do we go as far as registering individual tables and images, or do we keep the registry contents at a higher level—registering, for example, only the resources containing data collections. The choice of granularity determines the level of detail that can be incorporated into queries that locate resources in a registry, and what details are saved for queries to the specific data access services.

### 3.5 Metadata Standards / Relationships

No activities this year.

### 3.6 Metadata APIs

*API Specifications.* The SkyNode framework is a general approach to uniform querying of tables and data holdings. In addition to using the Astronomical Dataset Query Language (ADQL) as the query format (see WBS 3.4), it defines a set of service operations that allow different levels of support. These operations provide a framework for efficient joins across distributed tables. W. O'Mullane heads the development of the SkyNode specification within the context of the IVOA VOQL working group.

*Metadata Management: Prototype Service Directory.* This year, we established a special Registry "Tiger Team" to develop a prototype registry framework based on the model adopted by the IVOA Registry Working Group. It was completed to support NVO's Data Inventory Service (DIS) based on the Gamma-Ray Burst Science Demonstration. Components include:

- Publishing registries at NCSA and Caltech, where data providers can publish descriptions of resources including Cone Search and SIA services.
- A centralized, searchable registry at STScI that can harvest resource descriptions from the publishing registries using the OAI standard Protocol for Metadata Harvesting (PMH; <http://www.openarchives.org>).
- The Data Inventory Service portal at HEASARC: a user-oriented, web-based interface for locating data related to a position in the sky. It searches the STScI registry via a web service interface to discover the published Cone Search and SIA services; each of which is queried to find data associated with a user-provide position.

The DIS was released as the NVO's first end-user service in July at the IAU Assembly in Sydney.

As part of the registry development, R. Williamson and R. Plante have developed a VORegistry-in-a-Box package (<http://nvo.ncsa.uiuc.edu/VO/software>). This downloadable package allows a data provider to deploy a publishing registry on his/her own site. It provides forms for creating resource descriptions and exposes them to the VO using OAI. This package is being refined as the registry and metadata standards evolve.

Exchange of metadata between registries is a critical area of interoperability that we need to achieve on an international level. Although our use of OAI to collect metadata has been quite successful, other VO projects have not looked at this existing technology too closely as of yet.

## 4 Systems Architecture

### 4.1 System Design

The system design of the NVO architecture has the following components:

- Portals: Web Service interfaces to analysis procedures (OASIS, Aladin, the JPL YourSky, and the new Data Inventory Service from NASA Goddard.)

- Process management systems: Data processing pipelines to create derived data products (Chimera, Montage)
- Web Services: Uniform capabilities provided across NVO catalogs and image archives (cone search, VOTable catalog query, simple image access)
- Data Access Layer: Management of methods on data encoding formats for access based on physical quantities (UCDs) and a Data Format Description Language (DFDL).
- Data grid: Management of distributed collections, provision of logical name space for global persistent identifiers, and support for remote proxies (SRB)
- Computational grid: Access to distributed compute resources (Globus toolkit)
- Persistent archives: Management of technology evolution (SRB)
- Astrophysics catalogs and image archives (SDSS, 2MASS, DPOSS, et al.)
- Persistent disk systems: Interactive access to sky survey image collections (Grid Bricks)
- High performance disk caches: High speed access for bulk data analysis (SAN)
- Compute platforms: NSF TeraGrid

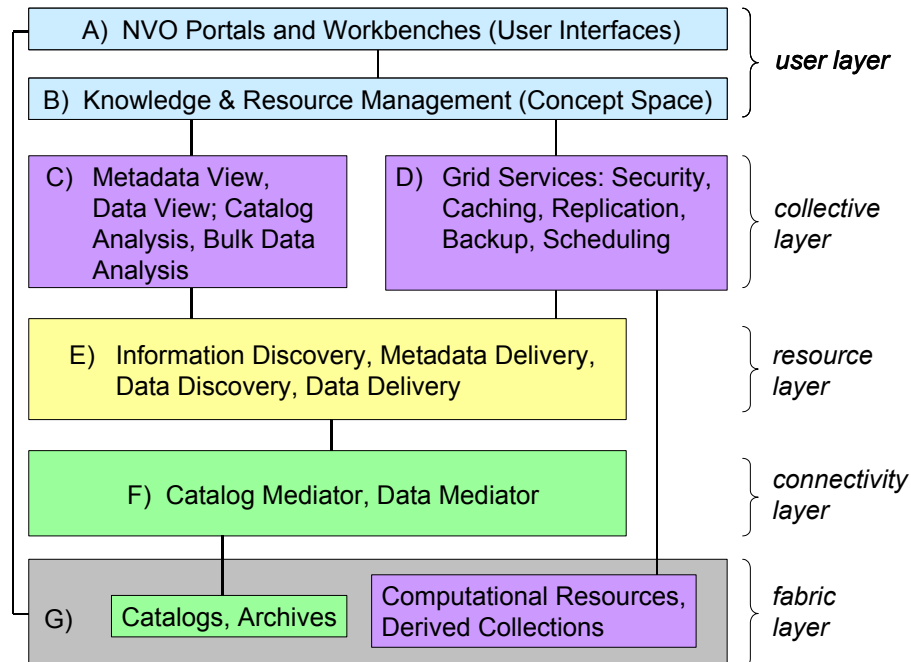


Fig. 1. The correspondence of the NVO architecture layers to the Grid infrastructure layers is shown on the right side of the diagram. Each component is designed to support access to the existing survey digital libraries and to the expanded capabilities required by the NVO to support analyses that require processing of a large fraction of the catalog holdings or images from multiple surveys.

The architecture specifies seven software layers and four resource layers (see Fig. 1). Components have now been developed for all of the layers. Since the design is based on loose integration of capabilities, it is possible for each higher software level to directly access the lower resource layers. The NVO architecture is roughly compatible with the

Grid architecture, with the top five levels being a refinement of the Grid application level. Much of the focus of the Grid architecture for operation across compute resources is subsumed in level six ('F', catalog/data mediator) of the NVO architecture.

The data access layer supports manipulation of digital entities based upon the semantic tags and knowledge relationships present within the digital entity. This in turn requires a data model, the ability to map to Unified Content Descriptors, and the ability to organize semantic terms relative to the NVO concept spaces for space, time, and domain knowledge. A simple catalog access method, the Cone Search protocol, and the Simple Image Access Protocol form the initial components of the data access layer. A Simple Spectral Access Protocol is in development. These DAL components rely on a well-specified data model, something that requires further development. The catalog and image access protocols could be defined because catalogs and images have quite intuitive data models. Once the data models are fleshed out, the Cone Search, SIA, and SSA protocols are likely to be updated.

A major issue is the appropriate mechanism for collections to be referenced from Grid technology. The NVO is currently supporting two implementations:

- File based access to images within sky surveys
- Collection based access to images within sky surveys

The SRB data grid provides collection-based access to 2MASS, DPOSS, and SDSS. The Chimera data processing system relies on file-based access.

A second issue is the distribution of analysis tasks between data management systems and Grid technology. The NVO demonstrations pointed out a continued need to refine the system design. A case in point is support for fine-grained (low-complexity) operations, versus large-grained (high-complexity) operations. Here complexity is measured as the number of floating point operations required per byte of data moved. Image sub-setting operations are typically low-complexity, and should be implemented directly at the storage resource within the data access system. High-complexity operations can be performed more efficiently by moving the data to a remote compute platform, and are excellent candidates for Grid workflow management systems such as Chimera.

#### *4.1.2 System-Level Requirements Definition and 4.1.3 Interaction with Grid Components and Tools*

There is a strong correlation between NVO system components and Research Groups within the Global Grid Forum. The NVO system design will need to track the results of the GGF working groups, in particular:

- The Data Format Description Language research group is developing standard XML-based description for characterizing scientific data. An XML file is created that identifies the structures within the digital data, the semantic labels applied to the structure, and the operations that can be performed upon the structures.
- The Grid File System research group is developing a standard for organizing the logical name space used to identify digital entities within a grid. The proposed

standard is consistent with the name space management of the Storage Resource Broker.

- The Data Transport research group is developing a specification for the standard operations supported at remote resources for data manipulation. A paper was written by R. Moore that defines the operations currently supported within the NVO and other Grid projects, and has been submitted to GGF.
- The Workflow Management research group / Portal research group is developing standards for managing data flow and control flow environments. This is one of the essential standards needed by the NVO to integrate processing pipelines with the Grid.
- The Open Grid Services Architecture and Open Grid Services Infrastructure working groups are developing standards for life cycle management of Grid services.
- The Data Access and Integration Services working group is developing a standard set of services for interacting with databases. An effort is underway to understand whether the SkyQuery service developed by A. Szalay can be implemented using OGSA-DAIS services. This project will strongly influence the implementation and design of the OGSA-DAIS interface.
- The Metadata Research Group is developing standard characterizations for database schema. This group would define schema representations for use by OGSA-DAIS and the Grid File System research groups. This group will also consider the implications of the Metadata Encoding and Transmission Standard schema for organizing administrative, descriptive, structural, and behavioral metadata.

A provisional charter has been created for an Astronomy Research Group within the Global Grid Forum. N. Walton (AstroGrid) and R. Moore have volunteered to lead the Research Group. The goal is to promote interactions between the IVOA and GGF. This includes providing input to the GGF on the requirements of the IVOA community for Grid and web services infrastructure, and providing evaluations of Grid performance and robustness. At the Global Grid Forum 9 meeting, a mapping was proposed between the IVOA assessments that are underway, and the provision of NVO requirements to the above working groups.

#### *4.1.4 Logical Name Space*

SDSC has been creating logical name spaces for NVO collections through their registration into the TeraGrid data grid. The GGF Metadata Research Group recognizes four types of identifiers:

1. Unique identifier, based on an OID or handle
2. Logical name, used to organize a digital entity within a collection
3. Descriptive metadata, used to support discovery independently of the unique identifier or logical name
4. Physical file name

The SRB supports all four forms of digital identification, with the additional naming conventions mapped onto the logical name space as metadata attributes. A key

requirement for NVO is consistency between these naming conventions. This can be cast as a decision to apply hard state management technologies to the mapping between these identifiers.

A discussion point is the manipulation of data based upon UCDs. The current hope within the Grid Forum is that the DFDL working group will provide mechanisms to support the mapping of semantic attributes onto structures within digital entities, and that the Metadata Research Group will provide tools to facilitate the mapping.

#### *4.2 Interface Definition*

An important issue with respect to the Grid community is the use of the Open Grid Services Architecture, and the underlying Open Grid Service Infrastructure for managing the life cycle of Grid services. A release was made in June of the OGSA infrastructure. However debates between the OGSA and Web Services Description Language (Semantic Web) communities are still in progress. NVO needs to track the discussions, and the proposed Astronomy Research Group provides a good way to do this.

OGSA based web services can be created on top of existing WSDL/SOAP based services. SDSC has implemented a WSDL/SOAP interface to the SRB data grid, and is now implementing an OGSA/OGSI compliant interface through the WSDL/SOAP mechanisms. The goal is to support access to the image archives registered into the NVO testbed through the latest Grid compliant interfaces.

#### *4.3 Network Requirements*

The movement of large image archives over networks is dependent upon use of parallel I/O and latency management mechanisms that minimize the number of messages. Within the TeraGrid, data rates of 250 MB/sec have been demonstrated between SDSC and NCSA, limited by the rate at which the local disk could receive the data. Collections of files were moved using 50 parallel I/O streams. With parallel file systems on the disk caches, substantially higher rates will be achievable, on the order of 1 GB/sec. SDSC has demonstrated rates of 3 GB/sec between the local TeraGrid node and a SAN array.

The management of latency requires support for bulk operations when moving small files (for instance the 2-MB 2MASS images). The files are aggregated into a container before movement, and unpacked when received. Bulk operations for registration (listing of the files in the logical name space), load (registration and import of the files) and unload (export of the files) have been demonstrated using the SRB. Data rates are 5 times faster using the bulk operations with parallel I/O.

#### *4.4 Computational Requirements*

The generation of a scientific quality mosaic requires the re-projection of every pixel. A single 2-MB image can take up to 2 minutes to re-project on a single processor of the TeraGrid. Reprocessing 5 million images on 2000 processors within the TeraGrid would

take 85 hours. This level of computation could be supported by the TeraGrid, but will require an allocation request. At the moment, 9 projects have requested time on the TeraGrid. The NVO project needs to request an allocation in the fall of 2003.

Cosmology simulations that are run on the TeraGrid, such as ENZO, generate tens of terabytes of simulation results. The comparison of observational data and simulation data will be important over the next year, and will require access to NVO archives.

#### *4.5 Security Requirements*

The TeraGrid has implemented the GSI security infrastructure, version 2.2. The infrastructure is continually being updated in response to identifier security holes.

## **5 Data Access/Resource Layer**

### *5.1 Resource and Information Discovery*

The registry has been an area of active development in the NVO and IVOA. The year has seen a prototype registry built to match the nascent Resource and Service Metadata (RSM) incarnated in XSD as VOResource. This helped to flush out some problems in both RSM and VOResource. There are now new versions of these standards that are expected to be endorsed by the IVOA. The prototypes will need to be modified to deal with the improved schema. Some work has been performed on a mapping of the XSD to a relational database schema.

Using SOAP for the searchable registry prototype seems to have proved a good choice. It has allowed for re-use of and easy integration of many tools, including the Data Inventory Service (DIS) and Mirage. Internationally, VO India will soon release VOPlot, talking also to the NVO registry prototype, and GAVO (German Astrophysical Virtual Observatory) also has a system talking to the prototype. This confirms the usefulness of having a prototype that is easily accessible and providing what VO developers need right now. Many papers will be presented at ADASS XIII (Strasbourg, October 2003) on the registry prototype work.

A general interface to the registry needs to be defined. One approach to this is to make it a basic SkyNode and reuse the work in that area to create the query interface to the registry. In NVO this seems highly acceptable, even, desirable, but in the international community there is much debate on the use of some XQuery-based language for this – even while adopting SkyNode protocol for catalogues. The registry interface is a major issue for resolution at the IVOA meeting in Strasbourg. Having a well-defined interface will allow experimentation with multiple registry technologies. One interesting project will be to implement our agreed-upon interface with the Grid-based MCS registry.

Although we believe the OAI protocol should form an important part of the registry, alone it is insufficient as an interface. It does allow a low price of entry for data providers to publish local registries, which may be harvested to complete queryable

registries. It seems that a SOAP/Web Services interface would be more appropriate (and probably an extension of SkyNode).

### *5.2 Data Access Mechanisms*

USC/ISI has continued the development of the Metadata Catalog Service (MCS.) The original service was based on web services. The new MCS under development is being ported to the OGSA-DAI technologies, which provide a Grid service interface to databases such as MySQL and Xindice.

MCS is a Metadata Service for data grids. MCS provides a mechanism for storing and accessing metadata, which is information that describes data files or data items. In particular, MCS aims to store information about logical files. A logical file uniquely identifies the content of a file. The Globus Replica Location Service (RLS) can be used to locate the physical instances of that file. Among the attributes of logical files are: file creator, creating timestamp, etc.

The Metadata Service allows users to query based on attributes of data rather than data names. In addition, the MCS provides management of logical collections and views of files. Collections are used to impose and authorization structure on groups of files. A collection can for example be build by a community. Views allow individual users to form a new grouping of data that is of interest to them. Views however, do not impose any authorization on the metadata access.

Containers are also supported in the design of the service. Containers enable the manipulation of small files or data items that need to be stored and handled together by the storage systems in order to provide performance.

MCS is also designed to be flexible and allows users to define their own attributes

### *5.3 Data Access Protocols and 5.4 Data Access Portals*

Development of the data access portal is proceeding on two fronts. Data access protocols define the interface between client applications and data access services. Data access frameworks provide a reference implementation of the data access protocols, including integration of legacy data analysis functions.

*Data Access Protocols.* Since the specification of the Simple Image Access (SIA) interface in October 2002 over thirty SIA service instances have been registered. SIA, along with the Cone Search service developed earlier, was used to implement several VO science applications that were demonstrated at the AAS in January 2003, and at the IAU in July 2003.

Further development of the Data Access Layer (DAL) protocols has since been moved to the DAL working group of the IVOA. The kick-off meeting of the IVOA DAL working group was held in Cambridge, UK, during the week of May 12, 2003. The goal of this first meeting was to agree on the scope of the data access layer, the working group goals, and the priorities for DAL standards development for the next year.

Specific working group agreements were reached in the following areas:

- Concept of DAL portal
- DAL scope and high level architecture
- Principal data types within the scope of the DAL
- Mapping of data types to access services
- Priorities for implementing the data access services
- Roadmap and priorities for the next year
- Enhancements to SIA V1.1

The principal classes of data to be supported by the VO data access layer include the following:

- *Source catalog.* This provides a simplified catalog query mechanism, e.g., for object catalog overlays on images, or for astrometric and photometric calibration of data. Basic catalog access services are required for data analysis, hence integrated access to catalog data is included within DAL. Large-scale statistical analysis, catalog cross matches, etc., are handled elsewhere within the VO.
- *Image.* This includes 2D sky projections, spectral data cubes, long slit spectra, and ultimately sparsely sampled images such as IFU data. VO will emphasize calibrated data but it may be desirable to be able to use the VO framework and services to access raw data as well.
- *ID spectrum and SEDs.* This is a special case of the more general NDIImage, provided for ease-of-access to most common spectral data.
- *Time series.* Light curves, variability data, etc. This category does not include synoptic image data, which is handled via the image interface instead.
- *Event and visibility data.* Although much access to radio and high energy data is via reference images, precise statistical analysis of event data, as well as imaging of both event and visibility data, requires access to (usually calibrated) event and visibility data.
- *Generic dataset.* Encompasses all types of data handled via the DAL. Used for basic data discovery, with subsequent access to data handled by the more specific DAL services.

Each principal type of data handled by the DAL has a corresponding data access service that is specific to and optimized for that particular class of data. Each type of data has a corresponding data model that is implemented by the service. Most data access in the VO is access to virtual data generated by the DAL services. Often the same data can be viewed via multiple services, e.g., synoptic or multiband imagery could be viewed as an image, as a spectrum or SED, or as a time series. Event and visibility data could be viewed as a table or as an image, spectrum, time series, and so forth, depending upon the capabilities provided by the service provider and the type of analysis being performed by the client.

The Cambridge meeting identified the top two priorities for near term DAL standards development to be further enhancement to the SIA interface, and addition of a new Simple Spectral Access (SSA) interface for access to 1D spectra and SEDs.

Priorities for enhancement to the SIA interface include provision for defining globally unique dataset identifiers (required to identify and manage data returned by the DAL services), development of component data models to allow better characterization of data quality (required for automated multiwavelength data analysis), and development of standards for serialization of data models and datasets in XML (required to return homogeneous datasets to client applications). All of these areas require coordinated development between DAL and the other NVO and IVOA working groups. Progress has been made on all of these but work is still ongoing.

As a first step toward developing the Simple Spectral Access (SSA) interface, a survey of spectral data providers and consumers (candidate client applications) has been conducted. A draft specification of a uniform data model for both 1D spectra and SEDs has been produced with the data models working group. Specification of the initial SSA interface will take place in late 2003.

*Data Access Frameworks.* The initial DAL protocols are simple enough to be implemented “from the ground up” by service providers and client applications. As VO data access becomes more sophisticated this will no longer be practical; it will become necessary to provide reference-grade software implementing the data access protocols, both on the server and the client side. Integration and reuse of legacy data analysis software with the VO is also a goal.

The data access portal will implement the VO data models and provide the framework “glue” needed to write distributed data analysis and data mining applications to access and operate upon astronomical data represented within NVO via the data models. This will include defining a framework within which data mediators can be implemented to map external data objects into the internal NVO data models, as well as server-side computation capabilities for data subsetting and filtering during application of the data mediator, and application of server-side analysis or transformation functions to “virtual data”—data objects in the internal data model format as generated by application of a data mediator.

As a first step towards construction of a scalable data access and analysis framework a whitepaper exploring the architectural issues has been written. The basic architecture is a component framework that wraps legacy (or new) data analysis code as components, defining a standard component-container interface. Components execute within a scalable, distributed computational framework. In principle a variety of frameworks will be able to execute the same shared components.

Current efforts are focused on technology research to survey all relevant technology and ultimately select the technology to be used to implement a reference-grade component framework.

## 6 NVO Services

### 6.1 Computational Services

Testing of the first public release of the Montage image mosaicing service has been completed. This release is designated version 1.7.2 and runs under Linux on single 32-bit processors. It will be released publicly on October 15<sup>th</sup>, 2003. We have made substantial progress on development of a prototype architecture that accepts a request for a 2MASS image mosaic through a web portal, submits the job to a computational grid at USC/ISI and U. Wisconsin, and then informs the user by email that the mosaic is ready for pickup. The architecture is sketched in Fig. 2.

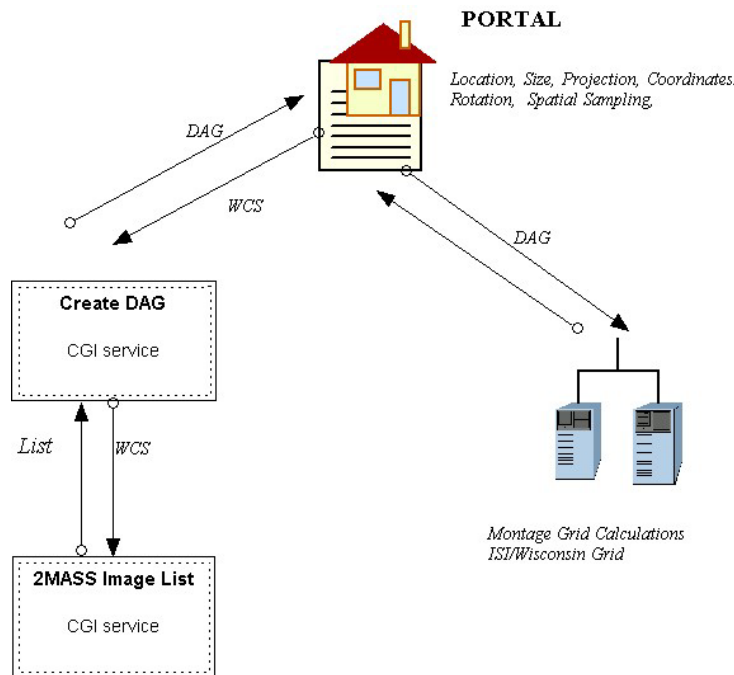


Fig. 2. Architecture of Montage as a Grid Service

The prototype will evolve into an operational service that will be deployed on the TeraGrid, when it is available. The architecture of Grid-enabled Montage that takes full advantage of the parallelization inherent in the Montage design and of software that runs jobs submitted to computing grids.

The heart of Grid-enabled Montage is a service that generates a Directed Acyclic Graph (DAG), which describes the processing flow in XML format, and illustrated in schematic form in Fig. 3.

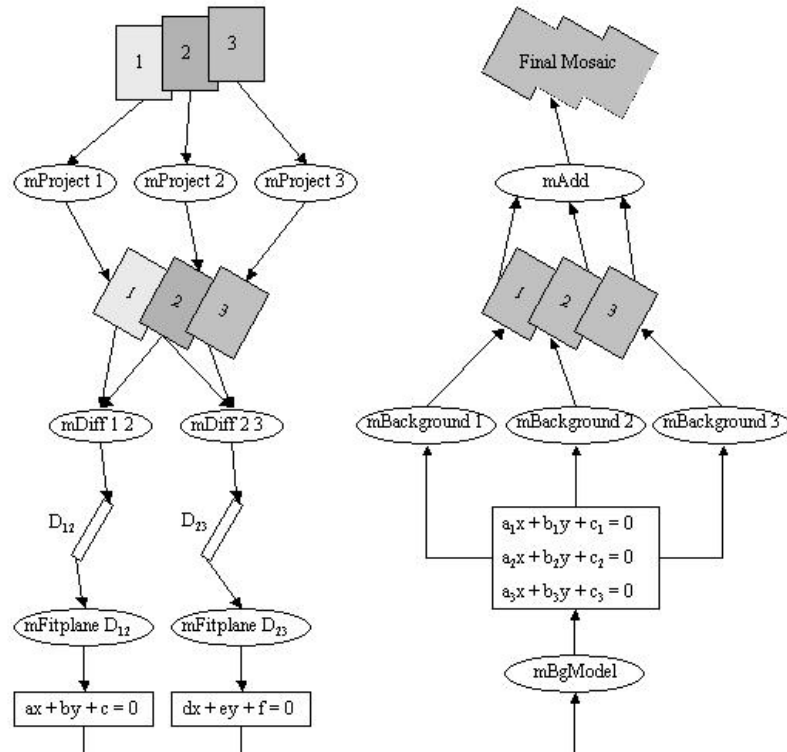


Fig. 3. Schematic of process and data flow in Montage.

This DAG is then passed to Pegasus, which generates a concrete DAG to be executed on a condor pool or a machine running PBS. Currently we are running montage over a single TeraGrid node or multiple condor pools. This concrete DAG identifies the path of the montage executables on the particular pool and adds nodes for transferring the images files from the GridFTP URLs to the execution directory. Once the final mosaic is generated it is also registered in the RLS. Currently we are not registering the intermediate data products. They can also be made permanent and registered in the RLS by modifying their persistence attribute in the abstract DAG. The Pegasus system was stress tested with the Montage code and some bugs were fixed in the Pegasus' code.

The projection jobs were made parallel in the DAG. There were 47 projection jobs in one of the test run. The scheduling of jobs on various machines was done by Condor matchmaker. All the projection jobs were completed in about 20 minutes whereas it takes about 2 to 2.5 minutes for a single projection on a standard Linux machine having a Pentium 4 processor running at 2 GHz. This shows the excellent speedup achieved by making the projection jobs parallel. Calculating difference images, etc., can also be made parallel using a similar mechanism. It took about half an hour in all to execute the concrete DAG and generate the final mosaic.

Unique names are provided for the image files. These images can be used across different computations since they are registered in the Replica Location Service. However we would need a unique naming scheme for the intermediate data products and the final mosaic also in order to make them available for future runs and discriminate products generated by different runs.

Once the compute infrastructure is in place, we will begin formal regression testing that will compare results on the 64-bit grid machines with those on single processor 32-bit machines.

## *6.2 Computational Resource Management*

USC/ISI evaluated the possibility of using the GriPhyN Virtual Data System to provide the necessary infrastructure for the execution of the galaxy morphology code in the Grid environment. We set up a Condor pool at ISI to perform the initial testing. Then we ran the galaxy morphology code on that pool. During our experimentation, we found that the code is dependent on particular versions of the OS. As the result, we modified the ISI computational pool to include only machines that can execute the analysis. We then expressed the galaxy morphology calculation as Virtual Data Language (VDL) constructs. This allowed us to have the computations automatically executed by VDS. VDS is composed of two parts: Chimera and Pegasus. Chimera generates an abstract workflow of how to produce the desired data and Pegasus maps the abstract workflow to a concrete workflow that can be executed on the Grid.

## **7 Service/Data Provider Implementation and Integration**

### *7.1 Service/Data Provider Implementation*

NVO organizations have implemented compliant Cone Search and Simple Image Access services in sufficient numbers as to enable all science demonstration projects and the project's first general science application, the Data Inventory Service (see WBS 8.1). These services include the following:

Caltech	DPOSS images, Messier catalog, Yale Bright Star Catalog
CADC/CVO	CNOC clusters of galaxies
HEASARC	SkyView images; Einstein, EXOSAT, ROSAT, OSSE archives; FIRST, SUMSS, and WENSS radio surveys

IRSA	2MASS catalog and images
JHU/FNAL	SDSS catalog and images
NCSA	Astronomical Digital Image Library
NED	NED catalogs
NOAO	NOAO Science Archive, Landolt UBVRI photometric standards, USNO A2.0
SAO	Chandra data archive
STScI	MAST and HST archive holdings, GSC I and GSC II catalogs and DSS I and DSS II images, Hipparcos catalog

## 7.2 Service/Data Provider Integration

The catalog and image access services implemented in WBS 7.1 are indexed in a prototype resource registry (WBS 5.1) and accessed through a common user interface via the Data Inventory Service (WBS 8.1).

## 8 Portals and Workbenches

### 8.1 Data Location Services

The Data Inventory Service was released as a generally available NVO service. This service may be accessed at

<http://www.us-vo.org/data-inventory>

or

<http://heasarc.gsfc.nasa.gov/vo>

Continued refinement of this service has taken place but these are primarily bug fixes or modest changes to the user interface. The basic relationship between the DIS service and the registry has proven to be sound and stable. Additional disk resources have been devoted to the DIS cache. Although it is not linked to any major astronomy Web sites, the DIS service has had modest use by the community with approximately 15 regions searched per day in September.

Further development of the DIS service will be needed to appropriately filter services that are to be queried. The number of services returned to a user can already be substantial and may soon overwhelm the user if presented naively.

### 8.2 Cross-Correlation Services

See WBS 10.3.

### 8.3 Visualization Services

No work in this area.

### 8.4 Theoretical Models

See WBS 10.2.

## 9 Test-Bed

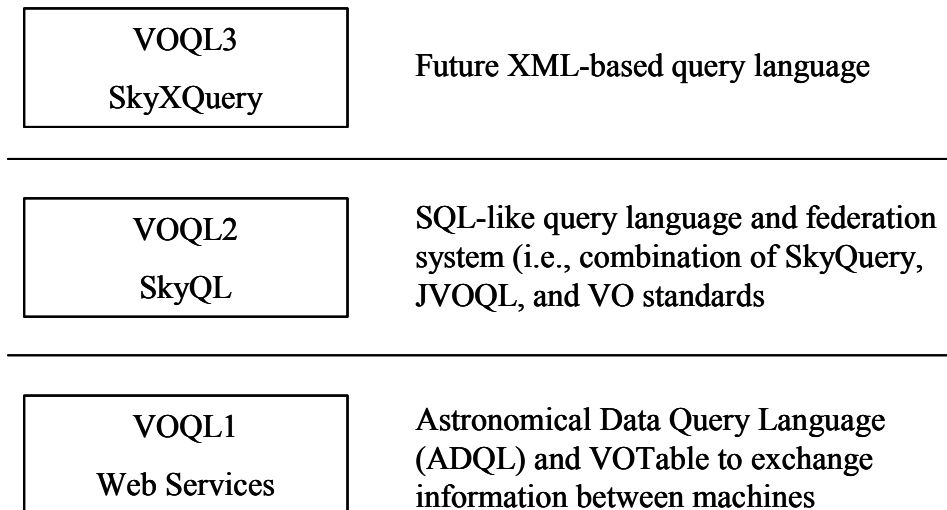
FNAL, USC/ISI and UIUC/NCSA have built a testbed that will support the NVO applications, in particular those that are being ported to the Chimera/Pegasus framework. Among such applications are the galaxy morphology science demonstration and Montage. The issues that need to be addressed in the FNAL environment is the use of Kerberos certificates. ISI has provided a solution to the problem.

When the TeraGrid became available in the summer of 2003, ISI has applied for accounts and has been conducting testing of the basic functionality for submitting jobs and data transfers.

## 10 Science Prototypes

### 10.1 Definition of Essential Astronomical Services

A Virtual Observatory Query Language (VOQL) remains a long-term goal of VO. This has now been crystallized into 3 layers as depicted below.



Much work was done earlier in the year on defining ADQL or layer one. Next effort was put into an IVOA proposal for VOQL1 and VOQL2 termed Basic and Full SkyNodes. Much of the proposal was built on knowledge gleaned from the Sloan Digital Sky Survey's SkyQuery ([www.skyquery.net](http://www.skyquery.net)) project. SkyQuery provided a glimpse of distributed astronomical queries. This system was built on SOAP Services but some Microsoft specific implementations were used. Now we wish to generalize this protocol and implement SkyQuery servers and portals on other platforms to demonstrate vendor-neutral interoperability. We are also extending the SkyQuery language and metadata to cover many more concepts and are harmonizing it with the Unified Content

Descriptors (UCDs) and with VOTable. There is considerable interest in this both within NVO and in the international community. The IVOA SkyNode proposal is backed by a diverse international group (<http://www.ivoa.net/twiki/bin/view/IVOA/IvoaVOQL>). We work most closely with the Japan VO project (Ohishi Masatoshi and Naoki Yasuda). A WSDL definition and basic implantation will soon be available.

General spectrum and filter access services have been made available at JHU as SOAP services. A web site has been built on top of these services allowing querying and upload of filter profiles and spectra. This work will fold into the SSAP (Simple Spectral Access Protocol) work (<http://skyservice.pha.jhu.edu/devel/wave>).

Extensive work has been done in making HTM work with general regions. This should be generalized to VO footprint and coverage services, i.e., intersection of regions for queries. These type of services are assumed in the SkyNode definition. Work still needs to be done on generic Web Services interface calls, in particular on the registry interface.

### *10.2 Definition of Representative Query Cases*

The continuing development and refinement of NVO services and capabilities must be guided by their scientific utility. To facilitate this objective there is a continuing need for the implementation of representative query cases and their presentation to the astronomical community. This process is essential not only to ensure the scientific quality of NVO activities but also to maintain an ever-increasing awareness of, and utilization of, the NVO by the US astronomical community. In addition, as VO activities in other countries become more widespread and mature, expanded efforts to increase communication and cooperation at an international level are critical in the joint development of query cases and scientific demonstrations.

Previous reports have described the creation of an NVO Science Working Group and the efforts of the group in assembling a preliminary list of 13 potential science query cases. These were further refined in 2002 to a list of three science demonstration projects to be developed during the year. These three science demonstrations were: 1) a brown dwarf search project; 2) a gamma ray burst follow-up service; and 3) a galaxy cluster morphology and evolution survey project. By the beginning of FY2003 these three projects were in the late stages of development, with the gamma ray burst demo nearing completion. The most complicated demonstration, and the one that involved use of Grid architecture, was the galaxy morphology demonstration, and intensive effort was put into this project in order to have it completed on schedule. Significant efforts during the last months of calendar 2002 by all the demonstration project teams resulted in all three of the demos being ready for presentation to the astronomical community at the January 2003 AAS meeting in Seattle. The general view of those involved in presenting the demonstrations is that they were very well received at this meeting. In addition, the brown dwarf search project discovered a previously unknown brown dwarf candidate, and confirming spectroscopy was obtained shortly thereafter. Thus the NVO enabled a new science discovery during the first few months after the implementation of these

scientific demonstrations. These demonstrations were also successfully shown at the IAU General Assembly in Sydney, Australia in June 2003.

Essential features of these science demonstrations have since been incorporated into on-line or forthcoming NVO services, as is described elsewhere in this report. Current activity in the area of query demonstrations is now focused on the integration of theoretical astrophysics and observational data through the mediation of NVO services and capabilities. This integration of theory and observation is in part an outgrowth of the Theory Virtual Observatory (TVO) effort described in last year's report. The particular example chosen for demonstrating the theory-observation interface is the evolution of globular clusters. This example was chosen because of the existence of a rich body of observational data and the extensive theoretical work done to date, and also because the topic is one of interest to a broad spectrum of astronomers in the subfields of stellar evolution, stellar dynamics, binary star evolution, and x-ray astronomy. The theoretical simulations can follow the dynamics and evolution of every star in a cluster of 100 million stars, including the formation and evolution of binary stars and the effects of stellar collisions and coalescence. Thus the calculations can produce color-magnitude diagrams (CMDs) as a function of time as well as profiles of stellar color and density as a function of radius and time. These theoretical datasets can then be compared directly with observed CMDs and image data for a series of different globular clusters. This comparison can then provide direct information for the first time about the history and dynamical state of a variety of observed globular clusters. The intent of the NVO Executive Committee and the globular cluster development group is to have this science demonstration available for presentation at the January 2004 AAS meeting in Atlanta.

In addition to the development of these science query cases, the NVO is promoting coordination with international VO partners on the design and development of other science demonstrations. To encourage this cooperation, the US NVO Project Scientist has become a member of the AVO Science Working Group, and other members of the US NVO team are also involved in discussions with their IVOA colleagues on these issues. The current AVO demonstration being developed will focus on spectroscopy and not on theory, in part as a complement to the US NVO activity. In addition, at the recent ADASS meeting in October 2003, an international theory VO collaboration was established among several IVOA members in order to facilitate the cooperative incorporation of theory datasets and tools into the international VO effort.

### *10.3 Design, Definition, and Demonstration of Science Capabilities*

*Gamma-Ray Burst Follow-Up Service and Data Inventory Service.* During late 2002, the HEASARC continued to lead the development of the gamma-ray burst follow-up demonstration. The demo showed how using VO protocols users could get and organize information from many different VO resources within just a few minutes of a burst. A fully functional gamma-ray burst demonstration was given at the January AAS meeting in Seattle, Washington.

As we developed the demo, it became clear that this service would also be useful as a

general tool for understanding what was known about any region specified by the user even when no dramatic event had occurred there recently. The burst demonstration was generalized into the VO Data Inventory Service. The simplified interface for this service was completed in early Spring 2003.

During the spring and early summer, in collaboration with the data registry activities at the STScI, the Data Inventory Service was modified to use dynamic VO registries. A fully operational Data Inventory Service was released at the IAU meetings in July 2003 and is now operational at the HEASARC and open to use by the professional and public communities. Currently approximately 80 image and catalog services from a dozen distinct institutions are queried by the DIS service. New services that are registered into the VO service registry at ST (or at one of the service registries from which the ST registry harvests data) are automatically added to the service.

Further augmentation to the DIS service is expected in the coming year, notably significant enhancements in how SIA services are supported, links to data archives as a standard for data set identifiers emerges, and support for more and new kinds of services, notably the Simple Spectral Access services.

*Galaxy Morphology Analysis.* US/ISI developed a Web Service front-end and the Grid processing engine for the galaxy morphological computations. The benefits of a Web Service front-end are that it is platform and language independent. Also it can be evolved into an OGSA compliant Grid Service. Specifically the Web Service server was running on a Linux box with Java implementation of the service while the client at the portal was running on a Microsoft .NET platform with C# implementation of the client. The Web Service provided an operation to pass an input VOTable serialized as a string and the name of the output VOTable to the application. The output of the operation was an HTTP URL where the status messages regarding the computations can be provided.

The Web Service creates the VDL constructs for the galaxy morphological computations from the input VOTable and submits the VDL to the Chimera/Pegasus for execution over the Grid. Each input VOTable can consist of hundreds of galaxy descriptions. The operation of the Web Service is to

- Create the VDL construct from the input VOTable
- Extract the HTTP URLs for getting the FITS image files for each galaxy
- Check if the image file is cached at the Web Server
- If the image files are not cached at the Web Server then the server downloads them and caches them at the server and also registers these individual image files in the Replica Location Service so that it can be found by Pegasus. Note that these images files are input to the galaxy morphological computations.

Pegasus registers the results of individual computations in the Replica Location Service so that they need not be computed again. Also the output VOTable is also registered in the RLS for the same reason. The uniqueness of these data products is guaranteed by using unique names for them.

a) Adding code to create a VOTable as the result of a galaxy morphology computation

Each input VOTable can have hundreds of individual galaxies. The galaxy morphological code has to run to each of these individual galaxies and the results have to be aggregated in an output VOTable. Each galaxy morphological computation creates a record that has to be aggregated into a output VOTable. We have written a program to create a final VOTable based on predetermined headers and footers and the output records from the individual computations.

#### b) Providing status information to the portal

The output from the Web Service is an HTTP URL where the status messages can be provided. The status messages can be provided up to any level of detail. It can show the status of the individual image files (downloaded or pending to be downloaded) and the status of individual galaxy morphological computations for a particular input VOTable. When the final output VOTable has been created the URL shows a status message such as:

Job Completed Final VOTable : <http://server.isi.edu/nvo/run/120909809/out.txt>

It provides the http URL from where the output VOTable can be retrieved. Currently only the final Job Completed message is displayed since it is more convenient for the portal.

The URL returned by the Web Service is implemented using a Java servlet that queries the status of individual computation from the Replica Location Service and provides the output messages. Every time the http URL is accessed, the servlet accesses the RLS to provide the most current status. The interval and frequency of polling the URL can be determined by the portal based on the size of the input VOTable.

*Brown dwarf candidate search.* A cross-comparison engine underwritten by NPACI was used to perform a brown dwarf candidate search by cross-comparing the 2MASS and SDSS catalogs. Out of the 326,020 cross-correlated pairs in the 2MASS IDR2 and SDSS EDR (~150 sq deg), we found 2,889 objects with  $z-J > 2.75$ . These are possible late-L to T dwarf candidates. Although all 2,889 candidates were not scrutinized, we culled out the best of these based on the closeness of the match between the two surveys (i.e.,  $< 1.0''$ ) and  $i-z$  colors of  $> 1.0$ . Our candidate list consisted of seven good candidates. Two of these were previously known: one T dwarf and one mid-L dwarf. Two of the others appear not to be real and are likely spurious sources matched in both surveys.

The final list of five candidates is as follows:

2MASSI J0016084-004301	suspected L dwarf
2MASSI J0104075-005328	<i>confirmed as mid-L dwarf (Keck LRIS spectrum)</i>
2MASSI J0229279-005328	suspected L (or T?) dwarf
2MASSI J1326298-003831	previously published L5
2MASSI J1346464-003150	previously published T6

## 11 Outreach and Education

### 11.1 Strategic Partnerships

An informal partnership has been formed with UC Berkeley, the American Museum of Natural History (Hayden Planetarium), and ManyOne Networks (<http://www.manyone.net>) to provide NVO content to the newly developed ManyOne web browser. ManyOne plans to demonstrate their new browser with some simple, 3D astronomical content (nearby stars) at the January 2004 AAS meeting.

A partnership was initiated with *Sky and Telescope* magazine to construct a VO-compliant archive of amateur astronomer images. Lack of personnel has temporarily side-lined this effort. We plan to restart this project in Year 3.

### 11.2 Education Initiatives

J. Raddick, the EPO lead for the Sloan Digital Sky Survey, presented a conceptual demo of an NVO-based curriculum unit at the January 2003 AAS meeting. The activity draws upon the SDSS and Chandra archives and enables students to “discover” quasars.

The Resource Metadata document (see WBS 3.4) incorporates metadata elements pertinent to education and outreach resources. Users will be able to search the resource registry for information based on intended audience (grade levels, for example), media format, and other characteristics (animation, artwork, photographic materials, press releases, etc.).

### 11.3 Outreach and Press Activities

A press release was issued on 12 March 2003 describing the brown dwarf science demonstration project and the newly discovered brown dwarf (“Virtual Observatory Prototype Produces Surprise Discovery,” *Headlines@Hopkins*, 12 March 2003, [http://www.jhu.edu/news\\_info/news/home03/mar03/nvo.html](http://www.jhu.edu/news_info/news/home03/mar03/nvo.html)). Follow-up articles appeared in Spaceflight Now, SpaceNews International, NPACI Online, UPI, and in *The New York Times*. A list of VO-related articles in the popular press is given in an appendix to this report.

## Activities by Organization

### Caltech–Astronomy Department

The Caltech NVO project is in close collaboration with the NFS Astrostatistics project at Penn State (G. Jogesh Babu, PI). Web services are being built as wrappers for these sophisticated statistical algorithms. These statistical algorithms include: K-means clustering, Multidimensional density estimation, and n-Point correlation functions. In the Virtual Observatory tutorial at ADASS Strasbourg, October 2003, there will be a session on the web services developed at Caltech.

### Caltech–Center for Advanced Computational Research (CACR)

Staff at CACR have contributed to the NVO development efforts in the following WBSs:

1. Management. Staff at CACR have been updating and maintaining NVO web pages, mailing lists, and discussion archives. There are regular new features about the status of the project.

2. Data Models.

(a) UCD Ontology. There is a new, international, UCD Steering Committee, chaired by R. Williams, to provide the balance between flexibility and interoperability. This committee provides the means for new UCDs to be added to the existing set.

Caltech has been leading (with CDS Strasbourg) the international discussion group on UCD (Unified Content Descriptor), an emerging shared semantic vocabulary for VO. The discussion serves as a foundation for the upcoming IVOA meeting in Strasbourg, where we expect to solidify the meaning and syntax of UCDs.

The phenomenological approach to metadata encapsulated in the UCDs has much similarity to the OWL ontologies being discussed in the Data Models working group. We expect to exploit this convergence next year.

The UCD structure is no longer a simple hierarchy of astronomical data terms. Now it is split into triples: [Concept, Property, Value]. Each part of the original tree can now play a role in one of the triples. For example, in the sentence "My apple weighs 400g", we have a Concept (apple), and an instance of the Concept (My apple), then the property (weight), and the value (400g).

(b) Future of UCD. We expect to make possible a new sophistication in confluent searches in this way. For example, if the request is for data that is "flux of polarized radio emission." we can use the new ontology to do better than a keyword search. It would "understand" that in the request above, the "polarization" idea and the "radio" idea must be closely linked, and that "flux" comes in many guises.

3.1 Metadata Standards. The Hyperatlas project has formed under NVO influence, a collaboration of Caltech, Jet Propulsion Lab, and San Diego Supercomputer Center,

defining standard WCS projections for resampled collections of astronomical images. A collection of such projections form the “pages” of an “atlas” that covers the sky in a uniform way, at a uniform scale. The project has created web services that give the standard projection corresponding to a given page number or point of the sky.

The intent of the Hyperatlas project is to encourage image resources to be projected to the same pixel grid, as specified by the Hyperatlas standard, so that sophisticated data mining algorithms can be brought to bear on the pixels. These algorithms include faint object detection in multiple wavebands, visualization of complex multi-wavelength imagery, and sophisticated color filtering. Over the next several months, the Caltech and SDSC groups will use the Montage software and the NSF TeraGrid to build standard atlases from some large sky surveys.

The Hyperatlas standard is define by a service that takes in an atlas specification; and either a page number in the atlas, or a position in the sky; the service returns the WCS part of the FITS header that for that page.

3.4 Profile Applications. Caltech worked closely with other NVO organizations—specifically NCSA and STScI—to implement the prototype NVO resource registry. Caltech created a local registry using the OAI (Open Archives Initiative) metadata harvesting protocol, the contents of which were ingested into the central NVO registry at STScI/JHU.

Caltech has worked closely with other NVO organizations—specifically NCSA and STScI—to bring up a working example of a distributed global registry of resources and services. Such a testbed has been created, based on the OAI (Open Archives Initiative) metadata harvesting protocol. At Caltech, two implementations of OAI have been build, one Java, the other Perl. The latter allows easy transformation between metadata schemas, and will be used to create a SIAP registry through self-publication. Registries and Caltech, NCSA, and STScI will harvest each other to make the distributed registry.

5.3 Data Access Protocols. At Caltech, SIAP services have been implemented for the 2MASS image survey, which resolves to the SRB Data Grid as well as a web server. The connectivity to the SRB system establishes a high-performance, high-bandwidth connection between NVO services and the NSF TeraGrid project.

The NVO Simple Image Access Protocol is being utilized in the Atlasmaker software, which acts as a harness for resampling codes such as NASA-funded Montage. SIAP is being used to transfer thousands of image files representing many terabytes of data to the NSF TeraGrid. Caltech has built a SIAP service for the pre-release 2MASS dataset, and utilized a SIAP service for SDSS images, so that we will compare and mine the federation of 2MASS and SDSS in the image domain. Atlasmaker has been built to work with two kinds of URL for fetching images, with the conventional `http://`, or the `srb://` protocol that uses the NSF-NPACI Storage Resource Broker Data Grid software to fetch images.

6.1 Computational Services. The Caltech group has been working closely with the NSF-TeraGrid project, doing large-scale image mosaicing with the Atlasmaker software. Atlasmaker uses Montage (described further below), a new and rigorous code for mosaicing images, as well as other, faster ways to build mosaics, for example for browse versions of images. Atlasmaker is one of the TeraGrid “flagship” applications. Under NPACI funds, the code has been parallelized and scripted for high-performance, wide-area computation on TeraGrid, using SRB for (some) input images, but essentially for the distributed storage of the resulting atlases. There is new code for connecting to arbitrary image archives that are using the NVO publishing protocol (SIAP). The protocol allows for multiple retrieval mechanisms: if the input data is on an SRB system, it can be retrieved that way, or else through HTTP. Code has also been built for the creation of atlases—coherent collections of mosaiced images that lead directly to multi-wavelength imagery. We expect these atlases to be a new and powerful paradigm for knowledge extraction in astronomy, as well as a magnificent way to build educational resources.

As the TeraGrid matures, we expect to be computing large numbers of mosaics, each a reprocessing of a particular image survey to a particular page from an atlas. The results will be stored back in a single virtual file system managed by SRB, but physically located at SDSC, JPL, and CACR.

#### **Caltech–Infrared Processing and Analysis Center (IPAC)**

The Infrared Science Archive (IRSA) has successfully evolved its architecture to support NVO protocols, and registered SIAP-compliant services that serve the following image collections housed at IPAC: IRSA Sky Survey, IRAS Galaxy Atlas, IRAS Mid Infrared Galaxy Atlas, IRAS Extended Galaxy Atlas, Midcourse Space Experiment, 2MASS All Sky Quicklook Images, Infrared Telescope in Space (IRTS), the ISO SWS Atlas, and the 2MASS 6x Lockman Hole Atlas. These images are served through a single application, and future image collections served through it will be automatically NVO-compliant.

IRSA has adopted the approach of developing a single application that responds according to the type of request made to it. For instance, if the request to the service is for HTML output, it will generate HTML that will be displayed in the client’s browser. If the request is for an inventory of a region, then the service will write information about the available data to a staging area. If the request is from an NVO-compliant program, then the service returns catalog information in VOTable format, or image information that is compliant with the SIAP.

IRSA’s VO service implementations show how NVO protocols can be supported with minimal modifications to an existing archive architecture. The NVO web services (catalog cone search and an image metadata search) both work via an HTTP Get, and can be supported by minor modifications to IRSA’s CGI services. The cone search is set up so the only free parameters are the CGI keywords `RA`, `DEC`, and `SR` (sky location in J2000 decimal degrees and search radius in degrees). The rest of the information (the “base” URL) must be fixed. For example, while IRSA has only one service that searches any of its INFORMIX catalogs, each one must be “registered” with a base URL like:

[http://irsa.ipac.caltech.edu/cgi-bin/Oasis/CatSearch/nph-catsearch?CAT=ntmass:ext\\_src\\_cat\\_01&](http://irsa.ipac.caltech.edu/cgi-bin/Oasis/CatSearch/nph-catsearch?CAT=ntmass:ext_src_cat_01&)

to which three parameters can be attached:

```
RA=12.8&DEC=-33.4&SR=0.5
```

Similar remarks apply to the SIAP services.

For any IRSA service that can return a table subset or image metadata list, there will be a special “raw data” mode that returns appropriate VOTables. For example, the Cone Search “services” IRSA provides are actually a slight reworking of the catalog access code developed for OASIS. Two things were added: An “NVO” flag was set if the CATALOG parameter was detected, in which case the RA, DEC, and SR parameters were looked for instead of the normal *locstr*, *etc.*; and if that flag was set the output table was run through *tbl2votable* and the XML/VOTable results were echoed straight back to *stdout* (as mime type *text/xml*).

Finally, IRSA developed a prototype service, *tbl2votable*, a C module used by the above services to convert column-delimited ASCII tables to VOTable format.

NED has delivered and registered an SIAP compliant image access service, underpinned by a new image metadata table applicable to all the images types served by NED. Technical and implementation details are as follows:

- While designing Simple Image Access Protocol (SIAP) to NED's Image Archive, layers were built between the NED archive and NVO services (due to the extreme heterogeneity of the imaging datasets in NED). This includes new database tables and software, supporting these particular types of data and access to them.
- The “FourCorners Table” (a metadata table with extracted World Coordinate System (WCS) keywords from the FITS headers and similar information for non-WCS compliant FITS and JPG images, provided by dozens of observatories and space missions, was deployed. In the case of the 2MASS All-Sky Survey, NED used the data in the IRSA archive and also table entries for their sky coverage. This is very valuable to have in NED because the 2MASS sources can be cross-correlated with other objects in NED. After setting the groundwork of flagging each image with a WCS quality flag, NED constructed the version of the metadata required to support SIAP queries.
- The search software makes use of the WCS metadata. It extracts from the metadata tables a listing of images in the NED archive for a particular query, filters it by various conditions (e.g., being fully WCS compliant; FITS, but not WCS compliant; e by wavelength (color), date, etc.). For JPG formatted images, a list is returned of URLs with abbreviated information about color, context of images (radio maps or contour diagrams, rare scanned images from old atlases, and so on).

In summary, Caltech/IRSA

- Delivered SIAP compliant versions of all image collections served through IPAC. These services have been registered with the NVO.
- Acted as technical lead on the brown dwarf pilot project, which led to the discovery of a new L4.5 brown dwarf.
- Began a collaboration with ISI to develop the infrastructure to support a Grid enabled version of Montage.
- Deployed a fully documented demonstration of ROME.
- Began planning for future development of ROME.
- Registered SIAP compliant NED image services, underpinned by a new metadata table that includes all of NED's image holdings. The architecture developed supports the richness and diversity of data served by NED or accessed through NED in a single service.
- Matured an architecture that ensures NVO compliance of IRSA services with minimal coding and modest maintenance costs.

### **Canadian Astronomy Data Centre/Canadian Virtual Observatory**

The CADC/CVO (an Associate Member of the NVO) continues its work on the CVO Data Query Prototype, the problem of the Data Archive/VO interface, massive database queries, and content generation to enable the global VO.

- The CVO Database Query Prototype has been publicly available since February 2003. The initial Sybase implementation has been replaced by a DB2 system on our parallel database system. Current multiwavelength content includes the WFPC2 Association stacks, the 2QZ spectra, and the ROSAT All-Sky Survey. Revision 2 of the system was released in July 2003 and was demonstrated at the IVOA booth at the Sydney IAU meeting.
- A parallel IBM BD2 system was received in crates April 1, 2003 and became publicly available July 15, 2003. This system is the most powerful database implementation in astronomy and consists of 18 processors and 256 disks with a capacity of 7 Terabytes. A full-time Database Administrator dedicated to the CVO work has hired April 2003. Tuning and optimization work continues as our data content grows.
- CVO staff have played key roles in the work on Data Models and VO Query Language working groups in Cambridge and Strasbourg. The CVO Data Query Prototype uses a generic multi-wavelength data model and a query model that allows simultaneous specifications of constraint sets on arbitrary types of astronomical observations and catalogs. This is the first implementation of such a system that we are aware of and it represents an instance of a solution to one of the core problems of the VO.
- The CVO continues to add detail and depth to its definition of the interface between an archive or project and the Virtual Observatory. Formal definitions of the boundary have been useful in determining a direct way for projects and archives to be VO-compliant without retrofitting against existing infrastructure. These definitions have been implemented and continue to evolve to accommodate more general datasets.

- The CADC/CVO have installed 60 Terabytes of online magnetic disk storage to accommodate new VO content and to facilitate distributed processing of the VO content of our system to enhance the value of our holdings.
- CVO continues to develop content for the VO such as WFPC2 Associations and CFHT12k mosaic camera images. These are processed in compliance with our data models and Archive/VO interface definitions.
- Collaborations with the U.S. NVO, the German Astrophysical Virtual Observatory, and the Australian VO continue to be key elements of the CVO development strategy. We have implemented a centralized metadata database for WFPC2 images, 2QZ spectra, and ROSAT All Sky Survey data to allow efficient processing of sophisticated queries while, at the same time, adopting distributed data access (transparent to the user) for actual observational data access. The system has been operational since July 2003 without any failures in our global data access model.
- The CADC hosted the U.S. NVO meeting September 3-4, 2003 at Herzberg Institute of Astrophysics in Victoria, British Columbia.

**Carnegie-Mellon University/University of Pittsburgh (CMU/UPitt)**

No report received.

**Fermi National Accelerator Laboratory (FNAL)**

Fermilab contributed in two areas this year: the galaxy-morphology demo and an NVO compliant interface to the SDSS first data release.

The galaxy-morphology demo (WBS 10.3.1) involved the largest effort and was developed in two stages. The first stage involved finishing a first version of the demo (which was started last year) for the AAS meeting in January. This work was done in conjunction with several other members of the NVO project. Contributions were made to refining the design of the demo, preparing a requirements document, and integrating the Grid computing component into the demo. J. Annis worked with USC/ISI (who developed the web portal for the Grid computing component of the demo), providing information on the interfaces used in the demo. V. Sekhri enabled TAM (a computing cluster at Fermilab) to act as a Grid node accessible by ISI. J. Annis also provided the code that computes the galaxy morphology parameters and package it for distribution to other Grid nodes. Finally he closed the loop to make sure that all inputs to the code were fetched using NVO services. In the second stage, V. Sekhri and J. Annis worked on developing an improved version of Grid component of the demo. This work involved creating a front-end galaxy morphology web service and integrating it with a Grid service that runs the GriPhyN virtual data toolkit. The galaxy morphology web service accepts a query much like an SIAP query and eventually returns a VOTable containing URLs to galaxy cutouts and galaxy morphology parameters. The cutouts are fetched from the SDSS SIAP service (described next). The morphology parameters are generated dynamically using Grid services and the Chimera virtual data application.

In the second area of work, V. Sekhri developed both SIAP and cone search interfaces to the imaging data in the SDSS DR1 data release. The SIAP interface returns URLs to image cutouts of objects in the object catalog—it does not make cutouts exactly at a user

specified position. A second service provides access to the FITS files of the cutouts themselves. The DR1 catalog covers 2099 square degrees and has 53 million objects, and the full dataset is 3 terabytes. These services relies on the backend SDSS data distribution infrastructure, which is subject to change in the future, and thus the services are still considered experimental. An interface to spectroscopy awaits clarification of the SIAP protocol. (WBS 7.2.1). The spectroscopic data are already online; just the SIAP interface needs to be implemented.

In related work not directly funded by NVO, V. Sekhri is working on the Virtual Organization (VO) project to provide a simple interface for users to authenticate themselves to gain access to Grid computing resources. Such an interface is needed by a wide range of distributed computing projects (iVDGL, EDG) and will be useful for integrating NVO with Grid computing resources (WBS 5.2.2). This project (which is largely for use by iVDGL) is expected to continue through December.

### **High Energy Astrophysics Science Archive Research Center (HEASARC)**

HEASARC personnel attended the Victoria team meeting and discussed and described the issues for further development of the Data inventory service. During the Victoria meeting the possible integration of the data set identifier that are proposed for use by NASA archives into the NVO identifier framework was extensively discussed.

Substantial activity was taken in support of the DIS. A number of small bug fixes and enhancements were made and the service was released to the public. The web interfaces were modified to ensure compliance with the Americans with Disabilities Act.

The primary thrust of activities in the HEASARC during the past quarter has been the development of metadata consistent with the emerging VO standards for all HEASARC Web resources. The completeness of UCDs for handling observations in an active archive environment was examined and a report discussing possible limitations was circulated. UCDs were assigned to several of the most popular HEASARC tables. Most recently the resource metadata structures for all HEASARC database tables was generated. These metadata are currently being reviewed by HEASARC scientists for consistency and completeness.

The HEASARC has begun to implement an OAI service to provide a 'publishing' only registry. This will be the mechanism through which the HEASARC provides information about available resources to the community. While the HEASARC may choose to provide a more capable registry in the longer term, it was felt to be a useful experiment to understand the effort that would be required, since this represents the minimal level at which a site can provide information to the VO while still directly controlling the publication process.

HEASARC personnel continued to take an active role in the development of metadata, data models and query languages through participation in telecons and discussion groups.

### **Johns Hopkins University**

W. O'Mullane and N. Li have put up the CAS Jobs batch processing site. This has support for local storage in the form of 'mydb' and may be generalized for VO use. See <http://skyservice.pha.jhu.edu/devel/casjobs/>

T. Budavari has made a CASService available, <http://skyservice.pha.jhu.edu/develop/vo/casclient.aspx>, and worked on the SIAP for SDSS.

W. O'Mullane has done much work on the Registry prototype. Most recently with G. Greene and R. Plante, VOResource compatibility with xsd.exe was improved and a relation schema mapping was worked on. See <http://sdssdbs1.stsci.edu/nvo/registry/>.

S. Carliles has integrated the registry prototype service and JAVOT and SAVOT with Mirage (a data analysis tool). A form in Mirage now allows the user to submit cone search requests to a Cone Search service. A list of Cone Search services is created by querying the registry. See <http://skyservice.pha.jhu.edu/develop/vo/mirage>.

V. Haridas, W. O'Mullane, and N. Li have been defining ADQL, producing the IVOA SkyNode proposal and SkyNode WSDL.

V. Haridas has made the FITSIO library available through C#.

G. Fekete, A. Szalay, and W. O'Mullane have worked on the HTM and region support.

M. Nieto has incorporated the SDSS DR1 image cutout and finding chart service into the recently announced public Data Release 1 (DR1) of the SDSS catalog data. See <http://skyserver.sdss.org/dr1/en/tools/chart/>.

G. Fekete is deploying VDT (virtual data toolkit) at JHU making us part of Condor/Grid at Fermilab.

W. O'Mullane and T. Budavari have done many demonstrations of Web Services and clients including DIME (<http://skyservice.pha.jhu.edu/develop/vo/imgcutoutclient.aspx>) . JHU Web Services are listed on <http://skyservice.pha.jhu.edu/develop/vo/>.

### **Microsoft Research**

J. Gray assisted in advanced database and web services design and implementation, and participated in NVO team meetings. Gray wrote a number of papers related to NVO and gave numerous presentations on NVO and the World Wide Telescope.

### **National Optical Astronomy Observatories (NOAO)**

During the past year NOAO staff have contributed to the development of the VO data model, and to the development of other VO standards, largely in the area of spectroscopy. NOAO staff also participated in the science demonstrations for the community, and in support of that activity have provided access to NOAO archive holdings through the Simple Image Access Protocol. NOAO staff have also been exploring a VO compute

model to support operations and services on data collections. The effort has focused on exploring an architecture which leverages current work on Grid-based compute models, but which separates the computation from the transient and persistent data stores (i.e., the data model) to allow dynamic optimization over the available resources. These concepts will be validated in the coming year with the construction of web services that will support the discovery, exploration, and retrieval of the MACHO synoptic time-domain dataset.

### **National Radio Astronomy Observatory (NRAO)**

NRAO has contributed primarily in the areas of data access portals (WBS 5.4), data models (WBS 2.1, 2.2) and system architecture (WBS 4.1, 4.2). D. Tody chairs the IVOA Data Access Layer (DAL) working group. Specific activities included:

- Participated in NVO system architecture development, particularly in the data access layer area.
- Further development of the Simple Image Access (SIA) interface, including support for SIA implementations and the NVO demos.
- Java servlet-based NVO cone search implementations for the NVSS, FIRST, WENSS radio source catalogs.
- Produced a whitepaper introducing the concept of a component framework for distributed and scalable multiwavelength astronomical data analysis.
- Initial work on the Simple Spectral Access (SSA) interface. Conducted a survey of spectral data providers and applications. Requirements and preliminary design of the SSA interface.

The initial version of the NRAO data archive was released for public testing on October 1 2003. This currently totals about 5 TB and includes the entire VLA data archive, plus all newly acquired data from the VLA, VLBA, and GBT telescopes.

### **Raytheon/ADC**

The Raytheon Technical Services Company (RTSC) provided support in the following activities:

*Project-wide.* RTSC staff participated in the NVO Project Team meetings and at the IVOA registry workshops in Cambridge, UK. Staff created, populated, and maintain a CVS repository for NVO software products, available at <http://nvo.gsfc.nasa.gov/viewcvs/viewcvs.cgi/>. Staff gave talks on the VO Query Language (VOQL), on a user interface to VOQL, on distributed data mining with the NVO, and on the CVS software repository. Staff also participated in on-line discussion groups, including several IVOA-sponsored e-mail discussion lists. RTSC staff have taken a leadership role as co-chair of the VOQL working group, formed out of the IVOA registry meeting.

*WBS 2: Data Models.* RTSC staff have worked on the VOQL and on the data model for the IVO “quantity” object. Staff have worked on the design of a constraint-based language for VOQL. Staff have learned Web Ontology Language (OWL) and how to use Protege Ontology Builder, and with this knowledge staff have developed a Units

Ontology for the NVO. Staff has begun an analysis of the Instrument Markup Language and SensorML in preparation for developing Astronomical Instrument Ontology.

*WBS 3: Metadata Standards.* RTSC staff continued to participate in and support the Metadata Working Group, including the weekly telecons, with particular emphasis on the NVO registries and the VOQL (VO Query Language). Staff provided contributions to the development of the Space-Time DTD and other XML standards for the NVO. Staff have also drafted a white paper on a Query Mediator, whose goal is to achieve a search for astronomical data across heterogeneously described datasets using XML web services. In support of this effort, over 10,000 lines of supporting PERL code were written as part of the Query Mediator activity.

*WBS 10: Science Prototypes.* RTSC staff provided detailed inputs, comments, and recommendations on the NVO science demos. RTSC provided support for IMPReSS, as part of the GRB Followup Service demo. RTSC staff are participating in the IVOA online discussions (RWP02) regarding the registry requirements that are needed to support science demos and science use cases. Under the auspices of other research funds, staff continues to investigate scientific data mining techniques with a long-term goal of applying these techniques within the NVO. Staff attended a Scientific Data Mining Workshop and the International SIAM data mining conference in May 2003, plus gave a talk at the SPIE data mining conference in April 2003.

### **San Diego Supercomputer Center**

SDSC continues support for the formation of an initial NVO testbed. The goal is to support large-scale analysis on replicas of collections that are located near the computational resources. The expectation is that consistency can be maintained across the replicated collections through use of the SRB data grid technology. Tasks that have been completed in the last three months include:

- Acquisition of 8 TB of disk to support a replica of the USNO-B catalog. We continue to coordinate with S. Levine (USNO) on the appropriate time to begin the data movement. A Grid Brick system has been implemented at an effective cost of \$3,500 per TB. The cost is now down to \$3000 per TB.
- The registration of the USNO-B catalog into a SRB data grid is now underway. Disks have been shipped to Flagstaff, where data will be loaded, and then transferred to SDSC for installation in a Grid Brick.
- Testing of the Mosaic technology developed at IPAC/Caltech. The software (developed by J. Good) was ported to the NSF TeraGrid, and was the first distributed application run on the TeraGrid. We continue to use the Mosaic service as a way to test the robustness of the NSF TeraGrid, and are coordinating with IPAC/Caltech on the development of large 2MASS mosaics.
- Implementation of a test version of the SDSS catalog. The goal is to understand the performance that can be obtained on TeraGrid resources in support of massive queries. Testing is in progress, and required the development of bit-manipulation operations for DB2. V. Nandigam has created the schema in DB2 and worked through the porting issues for the tables, etc. He's been working on porting the stored procedures and triggers, but has had some difficulty. This was mostly due to calls to

system tables for metadata about the tables that are not accessed in the same way in DB2. He has six of the stored procedures complete and is working on the remainder. The catalog has been implemented in DB2, on a 64-processor Sun server.

- Porting of 2MASS and DPOSS collections to Storage Resource Broker version 2.1. The new SRB 2.0 version supports parallel I/O, bulk data registration, direct tape access, GSI 2.2, and access controls on metadata. These features improve either the performance (data transport rate, data registration rate) or control.
- Initiated discussion with J. Smillie (ANU, Australia VO) on the registration of the MACHO image archive into the NVO data grid. This effort is being coordinated with NOAO. We have registered a small number of MACHO images into a SRB data grid.
- The SDSS DR1 data is being loaded into the SRB data grid, for high-speed access on the TeraGrid.
- Implementation of the USNO-B catalog. The USNO-B schema has been created and V. Nandigam has some test data loaded. Loading the rest of the catalog is in progress.

### **Smithsonian Astrophysical Observatory**

SAO led the Data Model design (WBS 2.1, 2.2) and the Metadata design (WBS 3.1) efforts.

#### Related Activities:

- Implemented a prototype and a definitive SIAP interface to the Chandra Data Archive.
- Heavily involved in the discussions about dataset identifiers, particularly in ensuring conformance between IVOA identifier standards and the identifiers developed under the auspices of the ADEC for information exchange between data centers and the ADS.
- Participated in the group that worked on the Galaxy Morphology Demo at the Seattle AAS meeting.

### **Space Telescope Science Institute**

Staff at Space Telescope actively participated in the coordinated development and demonstration of the Galaxy Morphology VO prototype, one of the three initial science applications designed with the NVO. In this project development a web portal was written using .NET technology that allowed a scientific user to investigate the morphology of galaxy clusters. STScI staff coordinated much of the effort for defining the prototype and all the required astronomical resources necessary to complete the analyses. In several cases VO services had to be evolved to the required level to provide the full functionality needed for the demo. Some of these areas include integration of the portal engine with web services to Grid computing, definition of standard VO Cone Search and SIAP services, VOTable parsing, and modification to Aladin visualization client such that catalog VOTables could be imported and overlaid with multiple FITS images.

Several VO meetings and technology demonstrations took place at STScI in a collaborative effort with The John Hopkins University (JHU) Sloan Digitized Sky Survey

(SDSS) project team for the purpose of exchanging technologies identified for VO application development. These included web services, image visualization services, specific project implementations of Cone Search, and SIAP services. The meetings are conducted in a regular manner and are planned throughout the year to come.

A prototype searchable VO Registry was developed at Space Telescope in collaboration with JHU SDSS. The registry was populated using the original VO Cone Search Registry and SIAP services based on the emerging standard metadata schema, VOResource. The backend was built on top of a SQL Server database using .NET and then mirrored between ST and JHU. Testing can be performed on either site with the production system deployed at STScI. In order to fully implement a VO Registry, OAI (Open Archives Initiative) Harvesting were developed to gather other remote VO resources from OAI repositories, i.e. local registries, from Caltech and NSCA. This capability was demonstrated at several meetings, showing how users could register resources at these remote sites and perform metadata queries on the ST VO Registry to retrieve them. Registry web services provide users with administrative tasks (load, edit, and remove resources) and a simple SQL query interface for retrieval of metadata elements contained in the VOResource schema. A Registry Query Form Builder JSP interface was developed to aid users in constructing queries to the registry. It allows the user to build a form based on the VOResource metadata elements/attributes and submit the resulting query to see the matching resources.

This registry was further enhanced by providing a suite of web services and web applications that were integrated with the Data Inventory Service (DIS) and Mirage clients. In order to work with DIS the registry population was augmented with a suite of resources from the HEASARC archive. The DIS application was demonstrated at the IAU meeting for users probing for astronomical resources for a given spatial region on the sky. DIS categorizes registry resources by specific metadata types (e.g., catalog, image archive, pointed observations) and provides interface capability to download resulting VO service datasets into OASIS and Aladin visualization clients.

The Mirage web client queries the registry and offers a list of Cone and SIAP Services and the ability to enter the RA, DEC, and SR parameters to send to the service(s). The VOTables returned by the Cone services are loaded as separate datasets in Mirage, and the image URLs returned by the SIAP service(s) are added to an option list in the astronomical imaging module. This capability provides VO resource data analysis and multi-parameter visualization.

In collaboration with NCSA and GAVO, a new database schema has been developed to match the newest version of VOResource. The registry updates are in progress to include resource metadata grouped into generic and specific service types tables.

Several new VO standard services have been publicly deployed from the ST MAST archive interface. MAST users can now obtain multiple HST instrument exposure field and pointing table information in VOTable formats. There are ongoing efforts at STScI

to bring more HST holdings into the VO framework, including HST preview images as an SIAP services and a GOODS data SIAP service.

The GALEX public archive at Space Telescope has developed VO Cone Search services to this new mission data set. VOTables can be accessed from the archive holdings although the proprietary period extends into the latter part of this year. A C# VOTable parser developed for the Galaxy Morphology demo was further enhanced with this archive development.

Staff at STScI continue to actively participate in the NVO metadata working group and also several of the IVOA interoperability groups.

### **United States Naval Observatory**

The USNO Flagstaff Archive server has been upgraded, and now has the capability to return catalog data in the XML/VOTable format defined by the VO organizations. We have made substantial modifications to the catalog extraction code that now allow the users much greater flexibility in customizing the catalog subsets that they can retrieve. Additional work on both hardware and software has been done to prepare the server to work as part of a virtual pipeline. Progress has been made organizing image datasets that are not currently, but will ultimately be made available through the USNOFS service. A stand-alone Cone Search code that reads the USNO-B1.0 catalog was written and sent to SDSC and to NCSA to help with access to the copy of USNO-B1 that was delivered previously.

### **University of Illinois-Urbana/Champaign/National Center for Supercomputer Applications (UIUC/NCSA)**

R. Plante continues to chair the weekly telecons of the Metadata Working Group. In the last year, major topics have included

- Resource metadata and identifiers
- Resource registries
- Data Inventory Service
- VOQL, ADQL, and SkyNode

R. Plante led the team that developed the Galaxy Morphology Science Demonstration, coordinating efforts at several NVO institutions:

- ISI: Grid-based analysis
- STScI: demonstration portal
- Fermilab: algorithm integration
- IPAC, CADC, CXC, HEASARC: data access services

R. Plante and E. Deelman (USC/ISI) will present the demonstration at Supercomputing 2003; the paper was accepted in the conference's peer-reviewed proceedings.

After the completion of the first year demonstrations, R. Plante and R. Williamson have concentrated on research on metadata and resource registries. This has focused on three key fronts:

1. *Resource Metadata Definitions.* R. Plante continues to contribute to the evolution of the Resource and Service Metadata document (RSM). R. Plante and R. Williamson, likewise, continue to refine the VOResource XML Schema accordingly. R. Plante leads the Resource Metadata Work Package of the IVOA Registry Working Group and is collaborating in the current review of the resource metadata model. Plante has begun work on a general style guide for metadata definitions in XML Schema.
2. *Resource Identifier Specification.* Through the IVOA Registry WG, R. Plante moderated the development of a specification for resource identifiers; he is currently editing a Working Draft for submission to the IVOA standards process.
3. *Registry Prototyping.* R. Plante and R. Williamson participated in the NVO Registry “Tiger Team”, aimed at providing a registry prototype for the Data Inventory Service. They developed the deployable, publishing registry package, VOResource-in-a-Box.

### **University of Pennsylvania**

The time domain poses interesting challenges to the virtual observatory. Repeated imaging surveys can gather enormous volumes of data at modest cost. The analysis of these data generally requires extended synoptic studies that span entire databases. New surveys are under development that will produce truly enormous databases in the coming decade (e.g. Pan-STARS and LSST).

At Penn we concentrate our efforts on applications that emphasize the time domain. These applications are chosen to cover almost all VO established standards and at the same time to be useful to the astrophysical community. To proceed with these applications we needed a real astrophysical dataset. The MACHO database was a natural choice because members of our group are affiliated to the original survey (full ownership), and because it is one of the prime examples that time is an important factor.

Main efforts:

- 1) WHAT: Curate a large database that supports all VO query protocols. HOW: Creating a SQL based MACHO database. The database contains light curves as well as metadata on the images. WHY: Make MACHO database available to the community through the VO. Test queries with restrictions on time of the observation, such as:
  - find objects where observation is taken between  $t_1$  and  $t_2$ .
  - return fluxes for certain band for time periods of 2 years
- 2) WHAT: Federating archives with metadata databases. HOW: The PENN MACHO DB does not contain the images however it contains all metadata as well as links to places that images are stored (Australia, San Diego etc). WHY: Implement Cone Search and SIAP.
- 3) WHAT: Implement a local registry and tools to harvest other registries. HOW: Implement the current registry standard with using web services (similar to the one in JH and Caltech). Write a VO cookbook on “make your data into an NVO database” and “make your application NVO available”. WHY. This is our

“market research”. What adoption rate do we get? What are the difficulties from the curator’s perspective?

- 4) WHAT: Combining NVO web service applications with federated data for solving a science problem. HOW: Calculate retrospectively the orbits of possible TNO, asteroids etc using orbit calculator and using VO query language to search for observations in the space/time domain of a particular magnitude.
  - find observations in this part of the sky taken in time interval  $t \rightarrow dt$  and of magnitude  $< \text{magnitude}_0$
  - find *moving* objects in this part of the sky taken in time  $t+dt$  and are of magnitude  $< \text{magnitude}_0$

These type of queries will test, clarify the standards for the VO query languages that need time constraints. The definition of a moving object may contain another VO web service that compares observations at different time. WHY: Demonstrate the usefulness of a federated data for astronomical questions. We can encourage/motivate other surveys to join VO based on the value demonstrated in this application (e.g. NEAT, SEKS). Prepare for future surveys.

- 5) WHAT: Mine the MACHO LC DB. Develop tools for mining the MACHO LC / MACHO metadata. Searching for clusters and in particular for outliers in a large dataset of light curves is of particular interest. The MACHO DB contains 60M objects. Mining big and complicated datasets makes this an interesting problem from the data mining perspective. Finding ways to use data mining tools on the VO platform is of a great challenge itself. However our goal is NOT to find general data mining tools or general approaches to mine over the VO but to solve the particular problem of finding outliers in a set of light curves using a method of a mixture of clustering and classification techniques.
- 6) WHAT: For the current effort of developing TAOS pipeline the Penn group is developing a mid-layer runtime platform that enables heterogeneous and distributed applications to be united into a single pipeline. We are exploring the idea to expand this to the VO and Grid environment. We are planning to add a visual layer, which will make the federation of different NVO applications and databases a simple task.

### **University of Southern California (USC/ISI)**

Tasks undertaken by ISI during the FY 2002/2003:

- Porting Montage to the Chimera/Pegasus framework. ISI ported the Montage application to the Chimera/Pegasus framework. As the results ISI was able to run a small sample Montage computation using a small number of Grid resources. Ongoing work is targeting the TeraGrid resources.
- Adding new features to Pegasus to support large-scale applications, such as those targeted by NVO. ISI added features to Pegasus that will enable to handle the large amount of individual files generated during the execution of applications such as Galaxy morphology and Montage.
- Continued evaluation of the Globus Replica Location Services, in particular in the context of the Pegasus system. ISI deployed the RLS and used it for the computations involving the Galaxy morphology and Montage applications.

- Deploying a Grid testbed across FNAL, ISI, NCSA and the TeraGrid. ISI is increasing the testbed that was used for the Galaxy morphology computation to include resources from FNAL and NCSA.
- Porting the Metadata Catalog Service (MCS) to the OGSA-DAI environment.

**University of Wisconsin**

No report received.

## Participant Report

Over 70 individuals are involved with work on the NVO project at 20 different organization and institutions. 17% of the participants in the project are female, and 13% are students. The total effort dedicated to the project is approximately 17 FTE, which includes in-kind contributions (at roughly the 15% level) from many organizations. Of the total effort, less than 1.5 FTE, or 9%, is associated with project management.

Name:	Charles Alcock
Role in project:	Leader, effort at Penn
Estimated time spent on project:	5%
Type of work done:	Discussions of temporal issues in NVO. Negotiated to move MACHO data to US.
Citizenship:	USA
Name:	Robyn Allsman
Organization:	NOAO
Role in project:	Software developer
Estimated time spent on project:	2%
Type of work done:	Assure compliance of NOAO archive with NVO standards. Develop NVO services for time-domain data sets (e.g., MACHO).
Country of citizenship:	USA
Name:	James Annis
Organization:	Fermilab
Role in Project:	NVO/GriPhyN/iVDGL interface.
Estimated Time Spent on Project:	8% (in-kind)
Type of Work Done:	Planning, software design, science prototype application development, staff supervision
Name:	David Archbell
Organization:	SDSC
Role in Project:	Catalog optimization
Estimated Time Spent on Project:	5%, funded by NPACI DTF
Type of Work Done:	Managed implementation of the SDSS catalog
Name:	John Benson
Organization:	NRAO
Role in Project:	NRAO archive, VO interfaces to NRAO archive
Estimated Time Spent on Project:	5%
Type of Work Done:	Participated in implementation of VO services (cone search) for NRAO archive.
Name:	Bruce Berriman
Organization:	IPAC, Caltech

Role in Project: Manager of IPAC NVO task, technical lead for  
 Brown Dwarf Search pilot project  
 Estimated Time Spent on Project: 15%  
 Type of Work Done: Managed work of IPAC staff supporting the NVO,  
 attended project meetings.  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Kirk Borne  
 Organization: George Mason University  
 Role in Project: Scientific Data Mining, Science Use Cases,  
 Public Outreach  
 Estimated Time Spent on Project: 10%  
 Type of Work Done: Scientific data mining research, led Raytheon team  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Lisa Brieger  
 Organization: SDSC  
 Role in Project: 2MASS re-projection project support.  
 Estimated Time Spent on Project: 5%, funded by NSF NPACI  
 Type of Work Done: Managed interactions with Caltech and IPAC for  
 2MASS analyses  
 Gender: Female  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Tamas Budavari  
 Organization: JHU  
 Role in Project: Science Software Developer  
 Estimated Time Spent on Project: 20%  
 Type of Work Done: Spectral/Filter Profile Site/ SIAP for SDSS,  
 Meetings/Reviews

Name: Samuel Carliles  
 Organization: JHU  
 Role in Project: Student  
 Estimated Time Spent on Project: 25%  
 Type of Work Done: Mirage/Registry integration

Name: Greg Chisholm

Organization: NOAO  
 Role in project: Software systems engineer  
 Estimated time spent on project: 10%  
 Type of work done: Developed straw-main computation framework for astrophysics aimed at supporting NVO data processing with Grid technologies  
 Country of citizenship: USA

Name: Joerg Colberg  
 Organization: University of Pittsburgh  
 Role in Project: Software developer  
 Estimated Time Spent on Project: 50%  
 Type of Work Done: Developed web services for SDSS, integrating data mining software  
 Country of Citizenship: Germany

Name: Alberto Conti  
 Organization: STScI  
 Role in Project: NVO Service Developer for GALEX archive  
 Estimated Time Spent on Project: 5%  
 Type of Work Done: Developed Cone Search services for GALEX public archive  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: Italy

Name: Tim Cornwell  
 Organization: NRAO  
 Role in Project: NRAO management, VO interfaces to AIPS++.  
 Estimated Time Spent on Project: 5%  
 Type of Work Done: Supervised and participated in interfacing of NVO services to AIPS++.

Name: Mark Cresitello-Dittmar  
 Organization: SAO  
 Role in Project: Analyst  
 Estimated Time Spent on Project: 12%  
 Type of Work Done: Internal CfA management and admin., DM design  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Ewa Deelman  
 Organization: USC/ISI

Role in Project:	Leading research in Grid-based computing
Estimated Time Spent on Project:	5%
Type of Work Done:	System Design: Pegasus and MCS
Name:	Dave De Young
Organization:	NOAO
Role in project:	Project Scientist
Estimated time spent on project:	15% (in-kind)
Type of work done:	Provide science oversight; select and guide science demonstration projects. Contribute to NVO Roadmap.
Country of citizenship:	USA
Name:	Janet DePonte Evans
Organization:	SAO
Role in Project:	Project management
Estimated Time Spent on Project:	5%
Type of Work Done:	SAO monthly and quarterly reports
Gender:	Female
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Laszlo Dobos
Organization:	Eotvos University, Budapest
Role in Project:	Student
Estimated Time Spent on Project:	10%
Type of Work Done:	Spectra and Filter profile site & Web Services
Name:	Ian Evans
Organization:	SAO
Role in Project:	Analyst
Estimated Time Spent on Project:	7%
Type of Work Done:	Work on VO data model, internal CfA meetings and discussions
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Giuseppina Fabbiano
Organization:	SAO
Role in Project:	PI/Coordinator of SAO efforts
Estimated Time Spent on Project:	10%
Type of Work Done:	Attending conferences, serving in committees (AVO SWG; ADEC), internal coordination

meetings and discussions at SAO.

Gender: Female  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: George Fekete  
 Organization: JHU  
 Role in Project: Software Developer  
 Estimated Time Spent on Project: 20%  
 Type of Work Done: HTM, Condor, iVDGL, VDT

Name: Mike Fitzpatrick  
 Organization: NOAO  
 Role in project: Software systems engineer  
 Estimated time spent on project: 10%  
 Type of work done: Developed tools and enhancements to IRAF core system to utilize NVO facilities. Attended NVO team meetings, contributed to technical discussions, implemented VO access to NOAO catalogs and archive image holdings, and developed prototype VOTool, a VOTable browsing GUI for IRAF.

Country of citizenship: USA

Name: Niall Gaffney  
 Organization: STScI  
 Role in Project: Technical developer  
 Estimated Time Spent on Project: 10%  
 Type of Work Done: Participant of MWG and NVO tech team meetings. Development of NVO Prototype including GalMorph Demo, Query From Builder for Searchable Registry. MAST.

Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Matthew Graham  
 Organization: Caltech Astronomy  
 Role in Project: Post-doctoral research associate  
 Estimated Time Spent on Project: 75% FTE  
 Type of Work Done: Software and architecture  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: UK

Name: Jim Gray  
Organization: Microsoft Research  
Role in Project: Database and web services design and implementation  
Estimated Time Spent on Project: 25%  
Type of Work Done: Attended meetings, helped with SkyServer and SkyQuery implementation, wrote papers  
Country of Citizenship: USA

Name: Gretchen Greene  
Organization: STScI  
Role in Project: Local NVO coordinator and technical lead  
Estimated Time Spent on Project: 40%  
Type of Work Done: Participant of MWG and NVO tech team meetings, coordinating local NVO meetings. Design development of NVO Prototypes including GalMorph Demo, Searchable Registry, OAI. Service developer for ST holdings: GSC, DSS, MAST.  
Gender: Female  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Robert Hanisch  
Organization: STScI  
Role in Project: Project Manager  
Estimated Time Spent on Project: 80%  
Type of Work Done: All aspects of project oversight (schedule, work packages, budgets, reports); leadership in IVOA Executive Committee and Working Groups; participation in Metadata Working Group, leading Resource Metadata development  
Gender: Male  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Vivek Haridas  
Organization: JHU  
Role in Project: Student  
Estimated Time Spent on Project: 35%  
Type of Work Done: ADQL, SkyNode, FITS library

Name: Michael Harris

Organization: SAO  
 Role in Project: Computer programmer  
 Estimated Time Spent on Project: 100%  
 Type of Work Done: SIAP libraries, client/server system for archive federation, SAO science prototype development  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Zhenping Huang  
 Organization: Raytheon Technical Services Company  
 Role in Project: XML Developer, Java Programmer  
 Estimated Time Spent on Project: 10%  
 Type of Work Done: Developed prototype software  
 Country of Citizenship: USA

Name: Steve Kent  
 Organization: Fermilab  
 Role in Project: Lead of Fermilab work  
 Estimated Time Spent on Project: 8% (in-kind)  
 Type of Work Done: Management

Name: Carl Kesselman  
 Organization: USC/ISI  
 Role in Project: USC/ISI PI  
 Estimated Time Spent on Project: 2%  
 Type of Work Done: Project leadership

Name: Mihseh Kong  
 Organization: IPAC, Caltech  
 Role in Project: Software Engineer  
 Estimated Time Spent on Project: 75%  
 Type of Work Done: Designed and developed ROME  
 Gender: Female  
 Race: Asian  
 Disabilities: None  
 Country of Citizenship: USA

Name: George Kremenek  
 Organization: SDSC  
 Role in Project: Archive replication  
 Estimated Time Spent on Project: 20%  
 Type of Work Done: Managed interactions with Caltech and IPAC. Replicated multiple sky surveys, supported cut-out services.

Gender: Male  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Jeongin Lee  
Organization: USRA at NASA/HEASARC  
Role in project: Primary HEASARC software developer.  
Estimated Time Spent on Project: 100% (since February 2003)  
Type of Work done: Attend telecons and informal meetings, wrote DIS service, Developed metadata for HEASARC resources

Name: Karen Levay  
Organization: STScI  
Role in Project: NVO Service Developer for ST MAST Holdings  
Estimated Time Spent on Project: 5%  
Type of Work Done: Developed Cone Search services for MAST HST Holdings

Gender: Female  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Stephen Levine  
Organization: USNO  
Role in Project: Interfaces USNO catalogs to NVO  
Estimated Time Spent on Project: 30% FTE  
Type of Work Done: Designed and wrote software to interface USNO digitized astronomical catalogs into NVO-compatible formats; provided data services to other NVO collaborators; attended meetings.

Gender: Male  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Stephen Lowe  
Organization: SAO  
Role in Project: Software developer  
Estimated Time Spent on Project: 100%  
Type of Work Done: DM design, participation in SIAP and SSAP development, SAO science prototypes

Gender: Male  
Race: Caucasian  
Disabilities: None

Country of Citizenship:	USA
Name:	Tanu Malik
Organization:	JHU
Role in Project:	Student
Estimated Time Spent on Project:	10%
Type of Work Done:	OpenSkyQuery, SkyQuery
Name:	Jonathan McDowell
Organization:	SAO
Role in Project:	Lead, Data Models Working Group
Estimated Time Spent on Project:	20%
Type of Work Done:	Leadership of IVOA DM working group, e-mail discussions, telecons, internal CfA meetings and discussions, attending NVO team meetings and conferences, analysis of data formats, co-authored documents on Quantity and Spectral data models
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Tom McGlynn
Organization:	NASA/GSFC
Role in project:	Lead at HEASARC, Metadata Working Group participation, DIS review.
Estimated Time Spent on Project:	15%
Type of Work Done:	Attended meetings and telecons, interacted over mailing lists. Developed software systems, managed other HEASARC actors.
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Reagan Moore
Organization:	SDSC
Role in Project:	Member of executive committee, lead on system architecture design, manager of SDSC activities
Estimated Time Spent on Project:	5%
Type of Work Done:	Attended GGF meetings, wrote concept papers, tracked technology
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA

Name: Viswanath Nandigam  
 Organization: SDSC  
 Role in Project: Student / post-graduation  
 Estimated Time Spent on Project: 50%  
 Type of Work Done: Implementing SDSS catalog and test querying  
 Gender: Male  
 Race: Asian  
 Disabilities: None

Name: Maria Nieto-Santisteban  
 Organization: JHU  
 Role in Project: Software Developer  
 Estimated Time Spent on Project: 40%  
 Type of Work Done: Image Cutout, Cone Search, Meetings/Reviews, Grid, Data Access for VO

Name: William O'Mullane  
 Organization: JHU  
 Role in Project: Software Developer, Group Leader, Designer  
 Estimated Time Spent on Project: 50%  
 Type of Work Done: SIAP, Registry, SkyNode, ADQL, JHU project oversight, meetings/reviews/telecons

Name: Olga Pevunova  
 Organization: IPAC, Caltech  
 Role in Project: Software Engineer  
 Estimated Time Spent on Project: 75%  
 Type of Work Done: NVO compliant NED services  
 Gender: Female  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Jeff Pier  
 Organization: USNO  
 Role in Project: Leads USNO NVO team  
 Estimated Time Spent on Project: 3% FTE  
 Type of Work Done: Managed interactions with USNO and NVO, attended meetings  
 Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Raymond Plante  
 Organization: NCSA/University of Illinois

Role in Project:	Senior Personnel; WBS lead
Estimated Time Spent on Project:	50%
Type of Work Done:	coordination of distributed collaborations, standards (document) development, software design and development
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Pavlos Protopapas
Role in project:	System design
Estimated time spent on project:	90%
Type of work done:	Design and implementation of architecture for time series analyses
Citizenship:	Cyprus
Name:	Norbert Purger
Organization:	Eotvos University, Budapest
Role in Project:	Student
Estimated Time Spent on Project:	10%
Type of Work Done:	Loading data for new skynodes
Name:	Jordan Raddick
Organization:	JHU
Role in Project:	Education and outreach support
Estimated Time Spent on Project:	10%
Type of Work Done:	SDSS/NVO integration
Name:	Srividya Rao
Organization:	USC/ISI
Role in Project:	Graduate Student
Estimated Time Spent on Project:	5%
Type of Work Done:	Developing algorithms for workflow scheduling
Name:	Arnold Rots
Organization:	SAO
Role in Project:	Lead for Space-Time Metadata
Estimated Time Spent on Project:	25%
Type of Work Done:	Development of space-time metadata definitions and schema, attended team meetings and Metadata WG telecons
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA

Name:	Vijay Sekhri
Organization:	Fermilab
Role in Project:	Software developer
Estimated Time Spent on Project:	30%
Type of Work Done:	Grid computing infrastructure, web service development, C/C++/Java coding
Name:	Dick Shaw
Organization:	NOAO
Role in project:	Manager of NOAO NVO activities.
Estimated time spent on project:	3%
Type of work done:	Attend NVO team meetings, align NOAO technical developments with NVO.
Country of citizenship:	USA
Name:	Ed Shaya
Organization:	Raytheon Technical Services Company
Role in Project:	XML Developer
Estimated Time Spent on Project:	20%
Type of Work Done:	Design data model and query language
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Gurmeet Singh
Organization:	USC/ISI
Role in Project:	Graduate Student
Estimated Time Spent on Project:	5%
Type of Work Done:	Developing MCS, working with NVO applications
Name:	Jaskaran Singh
Organization:	USC/ISI
Role in Project:	Graduate Student
Estimated Time Spent on Project:	5%
Type of Work Done:	Adapting Pegasus components to NVO
Name:	Frank Summers
Organization:	STScI
Role in Project:	Education and Outreach Coordinator
Estimated Time Spent on Project:	50% (since 1 July 2003)
Type of Work Done:	Plan EPO program, build EPO partnerships
Gender:	Male
Race:	Caucasian
Disabilities:	None

Country of Citizenship: USA

Name: Alex Szalay  
 Organization: JHU  
 Role in Project: Principle Investigator, Project Director  
 Estimated Time Spent on Project: 20%  
 Type of Work Done: Overall project vision, guidance, cone search registry, HTM regions

Name: Ani Thakar  
 Organization: JHU  
 Role in Project: Science Developer  
 Estimated Time Spent on Project: 10%  
 Type of Work Done: Meetings, reviews, VO data access

Name: Brian Thomas  
 Organization: Raytheon Technical Services Company  
 Role in Project: Design of meta-data, data model  
 Estimated Time Spent on Project: 20%  
 Type of Work Done: Wrote concept papers, tracked technology, wrote prototype software

Country of Citizenship: USA

Name: Randy Thompson  
 Organization: STScI  
 Role in Project: NVO Service Developer for ST MAST Holdings  
 Estimated Time Spent on Project: 10%  
 Type of Work Done: Developed Cone Search services for MAST HST Holdings

Gender: Male  
 Race: Caucasian  
 Disabilities: None  
 Country of Citizenship: USA

Name: Doug Tody  
 Organization: NRAO  
 Role in Project: Lead on NVO and IVOA data access layer, member of system architecture design team, responsible for NVO/VO activities at NRAO.  
 Estimated Time Spent on Project: 75%  
 Type of Work Done: Attended and chaired meetings, participated in working groups, specified interfaces, tracked technology, supervised NVO services implementation at NRAO.

Name: Frank Valdes

Organization:	NOAO
Role in project:	Data modeling for spectroscopy
Estimated time spent on project:	5%
Type of work done:	Wrote technical papers on VO data model and on the incorporation of spectra in the VO
Country of citizenship:	USA
Name:	Mark Voit
Organization:	STScI
Role in Project:	Education and Outreach Coordinator
Estimated Time Spent on Project:	50% (through 1 July 2003)
Type of Work Done:	Plan EPO program, build EPO partnerships, develop EPO metadata and use-case scenarios
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	USA
Name:	Antonio Volpicelli
Organization:	STScI
Role in Project:	NVO Service Developer for GALEX archive
Estimated Time Spent on Project:	10%
Type of Work Done:	Developed Cone Search services for GALEX public archive and C# VOTable parser
Gender:	Male
Race:	Caucasian
Disabilities:	None
Country of Citizenship:	Italy
Name:	Phillip Warner
Organization:	NOAO
Role in project:	Developer of NOAO-VO data services.
Estimated time spent on project:	5%
Type of work done:	Implemented VO access to NOAO catalogs and archive image holdings in support of Cone Search
Country of citizenship:	USA
Name:	Boyd Waters
Organization:	NRAO
Role in Project::	Web technology, service infrastructure
Estimated Time Spent on Project:	20%
Type of Work Done:	Participated in implementation of VO services (cone search) for NRAO archive, interfacing of AIPS++ resources to VO.

Name: Roy Williams  
Organization: Caltech CACR  
Role in Project: Co-PI  
Estimated Time Spent on Project: 50%  
Type of Work Done: Management, documentation, building code  
Gender: Male  
Race: Caucasian  
Disabilities: None  
Country of Citizenship: USA

Name: Ramon Williamson  
Organization: NCSA/University of Illinois  
Role in Project: Research Programmer  
Estimated Time Spent on Project: 100%  
Type of Work Done: software design and development  
Gender: Male  
Race: White  
Disabilities: None  
Country of Citizenship: USA

**Publications**

Berriman, G. B., Good, J. C., Curkendall, D. W., Jacob, J. C., Katz, D. S., Prince, T. A., & Williams, R. 2003, "Montage: An On-Demand Image Mosaic Service for the NVO," in ASP Conf. Ser., Vol. 295 *Astronomical Data Analysis Software and Systems XII*, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 343

Berriman, G. B., Kirkpatrick, J. D., Hanisch, R., Szalay, A., & Williams, R., "The NVO Brown Dwarf Search: Why Get Excited Over One Brown Dwarf?" Paper presented at JD08 "Large Telescopes and Virtual Observatories", IAU General Assembly, Sydney, July 2003.

Borne, K. D. 2003, "Distributed Data Mining in the National Virtual Observatory, SPIE Conference "Data Mining and Knowledge Discovery", Volume 5098, pp. 211-218.

Budavari, T., Malik, T., Szalay, A. S., Thakar, A. R., & Gray, J., 2003, "SkyQuery—A Prototype Distributed Query Web Service for the Virtual Observatory." ASP Conf. Ser. 295: *Astronomical Data Analysis Software and Systems XII*, 31.

Cresitello-Dittmar, M., Evans, J. DePonte, Evans, I., Harris, M., Lowe, S., McDowell, J. C., & Noble, M. S. 2003, "National Virtual Observatory Efforts at SAO," in ASP Conf. Ser., Vol. 295 *Astronomical Data Analysis Software and Systems XII*, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 65

Deelman, E., Plante, R., Kesselman, C., Singh, G., Su, M., Greene, G., Hanisch, R., Gaffney, N., Volpicelli, A., Budavari, T., Nieto-Santisteban, M., O'Mullane, W., Annis, J., Sekhri, V., Bohlender, D., McGlynn, T., Rots, A., & Pevunova, O. 2003, "Grid-Based Galaxy Morphology Analysis for the National Virtual Observatory," *Supercomputing 2003*, accepted. <http://www.sc-conference.org/sc2003/paperpdfs/pap282.pdf>

Djorgovski, S.G. 2002, "The Roles of Small Telescopes in a Virtual Observatory Environment", in *Small Telescopes in the New Millennium. I. Perceptions, Productivity, and Priorities*, Dordrecht:Kluwer, ed. T. Oswalt, p. 85

Djorgovski, S.G., Brunner, R., Mahabal, A., Williams, R., Granat, R., & Stolorz, P. 2002, "Challenges for Cluster Analysis in a Virtual Observatory", in *Statistical Challenges in Astronomy III*, eds. E. Feigelson & J. Babu, New York: Springer Verlag, p. 125

Dowler, P. D. 2003, "Data Management for the VO," in ASP Conf. Ser., Vol. 295 *Astronomical Data Analysis Software and Systems XII*, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 209

Good, J. C., Kong, M., & Berriman, G. B. 2003, "OASIS: A Data Fusion System Optimized for Access to Distributed Archives," in ASP Conf. Ser., Vol. 295 *Astronomical Data Analysis Software and Systems XII*, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 89

Gray, J., & Szalay, A., 2002, "The World Wide Telescope: An Archetype for Online Science," MSR TR 2002-75, pp 4, CACM, Vol. 45, No. 11, pp. 50-54, Nov. 2002.

Gray, J., Szalay, A.S., Thakar, A., Kunszt, P., Stoughton, C., Slutz, D., & Vandenberg, J., 2003, "Data Mining the SDSS SkyServer Database," Distributed Data & Structures 4: Records of the 4th International Meeting, pp 189-210 W. Litwin, G. Levy (eds.), Paris France March 2002, Carleton Scientific, ISBN 1-894145-13-5, also MSR-TR-2002-01, Jan. 2003.

Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C., & Vandenberg, J., 2002, "Online Scientific Data Curation, Publication, and Archiving." SPIE 4846, 103.

Hanisch, R.J., and the IVOA Interoperability Working Group & NVO Metadata Working Group, 2003, "Resource Metadata for the Virtual Observatory Version 1.0," IVOA Working Draft.

Hanisch, R.J., & Linde, A.E., 2003, "IVOA Document Standards Version 0.2," IVOA Working Draft.

Jacob, J., Brunner, R., Curkendall, D., Djorgovski, S.G., Good, J., Husman, L., Kremenek, G., & Mahabal, A. 2002, "YourSky: Rapid Desktop Access to Custom Sky Image Mosaics," Proc. SPIE 4846, 53.

Mahabal, A.A., Djorgovski, S. G., & Williams, R.E., "Topic Maps for Semantic Access to the Virtual Observatory," AAS Meeting 201, #09.02

Mahabal, A., Djorgovski, S.G., Williams, R., Brunner, R. 2002, "Topic Maps for Custom Viewing of Data," Proc. SPIE 4846, 65

McDowell, J., "A Virtual Astrophysics Library in the Virtual Observatory," AAS Meeting 201, #150.01

McDowell, J. C. 2003, "Small Theory Data in the Virtual Observatory," in ASP Conf. Ser., Vol. 295 Astronomical Data Analysis Software and Systems XII, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 61

McGlynn, T.A., & McDonald, L., "SkyView: Ten Years with the Virtual Telescope," AAS Meeting 201, #09.01

Mink, D. J. & Kurtz, M. J. 2003, "Federating Catalogs and Interfacing Them with Archives" A VO Prototype," in ASP Conf. Ser., Vol. 295 Astronomical Data Analysis Software and Systems XII, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 51

Moore, R., "Recommendation for Standard Operations at Remote Sites," submitted to Global Grid Forum, Sept. 2003.

Moore, R., & Baru, C., "Virtualization Services for Data Grids", Book chapter in "Grid Computing: Making the Global Infrastructure a Reality", John Wiley & Sons Ltd, 2003.

Plante, R., Linde, T., Williams, R., and Noddle, K. 2003, "IVOA Identifiers Version 0.2", IVOA Working Draft,  
<http://www.ivoa.net/Documents/WD/Identifiers/WD-Identifiers-20030930.html>

Rajasekar, A., Wan, M., Moore, R., Jagatheesan, A., & Kremenek, G., "Real Experiences with Data Grids—Case Studies in Using the SRB," International Symposium on High-Performance Computer Architecture, Kyushu, Japan, December 2002.

Rajasekar, A., Wan, M., Moore, R., Kremenek, G., & Guptil, T., "Data Grids, Collections, and Grid Bricks", Proceedings of the 20<sup>th</sup> IEEE Symposium on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies, San Diego, April 2003.

Szalay, A.S., Budavari, T., Malik, T., Gray, J., & Thakar, A.R., 2002, "Web Services for the Virtual Observatory," SPIE 4846, 124.

Thakar, A. R., Budavari, T., Malik, T., Szalay, A.S., Fekete, G., Nieto-Santisteban, M., Haridas, V., & Gray, J., "SkyQuery - A Prototype Distributed Query and Cross-Matching Web Service for the Virtual Observatory," AAS Meeting 201, #105.07

Thakar, A.R., Szalay, A.S., Kunszt, P.Z., & Gray, J., 2003, "The Sloan Digital Sky Survey Science Archive: Migrating a Multi-Terabyte Astronomical Archive from Object to Relational DBMS," Computing in Science and Engineering, V5.5, Sept 2003, IEEE Press. pp. 16-29.

Thakar, A. R., Szalay, A. S., VandenBerg, J. V., Gray, J., & Stoughton, A. S. 2003, "Data Organization in the SDSS Data Release I," in ASP Conf. Ser., Vol. 295 Astronomical Data Analysis Software and Systems XII, eds. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook (San Francisco: ASP), 217

Thomas, B., Shaya, E., & Cheung, C., "An Extensible Query Framework for the Virtual Observatory," AAS Meeting 201, #08.05

Valdes, F. G., 2003, "Incorporating Spectra in the Next Phase of the Virtual Observatory," <http://iraf.noao.edu/projects/vo/dal/specsiap.html>

Valdes, F. G., 2003, "A Virtual Observatory Data Model," <http://iraf.noao.edu/projects/vo/dal/datamodel.html>

### **Virtual Observatory Articles in the Popular and Technical Press**

“Telescopes of the World Unite! A Cosmic Database Emerges,” 20 May 2003, B. Schechter, *The New York Times*.

“Virtual Observatory Demo Produces Surprise Discovery,” SpaceFlight Now, 12 March 2003, <http://www.spaceflightnow.com/news/n0303/12virtual/>

“Virtual Observatory Discovers New Star,” United Press International, 12 March 2003, <http://www.upi.com/view.cfm?StoryID=20030312-054957-7150r>

“A Virtual Observatory for the Digital Universe,” *Astronomy and Geophysics* 44 (2) 2.04 [http://www.blackwellpublishing.com/products/journals/aag/AAG\\_April03/aag\\_44204.htm#seq2](http://www.blackwellpublishing.com/products/journals/aag/AAG_April03/aag_44204.htm#seq2)

“European Virtual Observatory One Step Nearer,” *Sky and Telescope*, 28 January 2003, [http://skyandtelescope.com/news/current/article\\_850\\_1.asp](http://skyandtelescope.com/news/current/article_850_1.asp)

“Web-Based Virtual Observatory Discovers Star in Trial Run,” *SpaceNews International*, 24 March 2003

“Seeing the Sky in a Whole New Way,” *Mercury*, March-April 2003, [http://www.astrosociety.org/pubs/mercury/32\\_02/nvo.html](http://www.astrosociety.org/pubs/mercury/32_02/nvo.html) (partial copy)

“New National Virtual Observatory,” *Starry Skies*, November 2002, <http://starryskies.com/articles/2002/11/nvo.html>

“NVO Prototype Produces Surprise Brown Dwarf Discovery, Infrastructure Includes SDSC Storage Resource Broker,” *NPACI Online*, 19 March 2003, [http://www.npaci.edu/online/v7.6/nvo\\_discovery.html](http://www.npaci.edu/online/v7.6/nvo_discovery.html)

**Acronyms**

AAS	American Astronomical Society
ADC	Astronomical Data Center
ADEC	Astrophysics Data Centers Executive Committee (NASA)
ADQL	Astronomical Data Query Language
AIPS++	Astronomical Image Processing System++ (NRAO)
API	Applications Programming Interface
AVO	Astrophysical Virtual Observatory
CACR	Center for Advanced Computational Research (Caltech)
CADC	Canadian Astronomy Data Centre
CDS	Centre de Données astronomiques de Strasbourg
CMU	Carnegie Mellon University
CXC	Chandra X-Ray Center
CY	calendar year
DAG	Directed Acyclic Graph
DAGMan	Directed Acyclic Graph Manager (Condor)
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Projects Agency
DIS	Data Inventory Service
DM	Data Model
DOE	Department of Energy
DPOSS	Digitized Palomar Observatory Sky Survey
DTD	Document Type Description
EPO	Education and Public Outreach
ESTO	Earth Science Technology Office (NASA)
ESTO-CT	ESTO Computational Technologies (NASA)
FIRST	Faint Images of the Radio Sky at Twenty Centimeters
FITS	Flexible Image Transport System
FNAL	Fermi National Accelerator Laboratory
FTP	File Transport Protocol
FY	fiscal year
GB	gigabyte
GLU	Générateur de Liens Uniformes (uniform link generator)
GRB	Gamma Ray Burst
GriPhyN	Grid Physics Network
HEASARC	High Energy Astrophysics Science Archive Center
HTTP	HyperText Transport Protocol
IPAC	Infrared Processing and Analysis Center (Caltech)
IRAF	Image Reduction and Analysis Facility (NOAO)
IRSA	Infrared Science Archive (IPAC)
ISI	Information Sciences Institute (USC)
ITWG	Information Technology Working Group (NASA data centers)
iVDGL	International Virtual Data Grid Laboratory
IVOA	International Virtual Observatory Alliance
JDBC	Java Data Base Connectivity (Sun, Inc., trademark)

JHU	The Johns Hopkins University
MAST	Multimission Archive at Space Telescope (STScI)
MB	megabyte
MOU	Memorandum of Understanding
MWG	Metadata Working Group
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputer Applications
NOAO	National Optical Astronomy Observatories
NPACI	National Partnership for Advanced Computational Infrastructure
NRAO	National Radio Astronomy Observatory
NSF	National Science Foundation
NVO	National Virtual Observatory
OAI	Open Archives Initiative
OASIS	On-line Archive Science Information Services (IRSA)
OGSA	Open Grid Services Architecture
OIL	Ontology Inference Layer
OWL	Web Ontology Language
PB	petabyte
PMH	Protocol for Metadata Harvesting (of OAI)
Q	quarter
QSO	Quasi-Stellar Object
RC	Replica Catalog
RDF	Resource Description Framework
RLS	Replica Location Service
ROME	Request Object Management Environment
SAO	Smithsonian Astrophysical Observatory
SAWG	Science Archives Working Group (NASA)
SAWG	System Architecture Working Group (this project)
SciDAC	Scientific Discovery through Advanced Computing (DOE)
SDSC	San Diego Supercomputer Center
SDSS	Sloan Digital Sky Survey
SDT	Science Definition Team
SIAP	Simple Image Access Protocol
SOAP	Simple Object Access Protocol
SRB	Storage Resource Broker
SSAP	Simple Spectral Access Protocol
STScI	Space Telescope Science Institute
SWG	Science Working Group
TB	terabyte
UCD	Unified Content Descriptor
USC	University of Southern California
UDDI	Universal Description, Discovery, and Integration
UIUC	University of Illinois Champaign-Urbana
USNO	United States Naval Observatory
USRA	Universities Space Research Association
VDL	Virtual Data Language

VDS	Virtual Data System
VO	Virtual Observatory
VO	Virtual Organization
VOQL	Virtual Observatory Query Language
WBS	Work Breakdown Structure
WSDL	Web Services Description Language
XML	Extensible Mark-up Language
2MASS	Two-Micron All Sky Survey

# The National Virtual Observatory

## An Information Technology Framework for Astronomy

### Funded by the National Science Foundation

#### Project Development Roadmap

The National Virtual Observatory (NVO) is creating an environment for astronomical research that will enable the execution of research projects whose scale and scope have not been possible previously. This capability will lead to an era of unprecedented astronomical discoveries. The NVO will establish this environment through the use of high-performance computing, large-scale databases, and web and grid services. The NVO will establish standards for data representation and services, and it will integrate resources (catalogs, image archives, and processing pipelines) with standard services (image, spectrum, and catalog access protocols) to provide an environment of unprecedented power and simplicity for carrying out scientific research.

A series of science prototypes will guide the development of infrastructure and help define the core toolkit. We will work closely with the communities of end-users and developers. This interaction will help to define the science and technical demonstrations, and will engage a new community of VO developers--just as the development of the WWW ignited a new wave of creativity in information services. These services, such as a sophisticated global resource registry for the publication, discovery, and use of NVO-compliant resources, will be the basis for ground-breaking astronomical research. Advanced information technologies such as virtual data, distributed computing, grid services, and large-scale database access will facilitate the construction of complex data analysis and visualization tools.

The NVO is an excellent vehicle for education and public outreach, both in astronomy and in the underlying IT. We will increase the amount of data in NVO-compliant repositories that is useful for outreach activities, and develop the software tools and services that can readily identify and access suitable NVO datasets. We will help create portals to NVO datasets to that are useful to outreach professionals, educators, and/or the general public. We will show how modern IT infrastructure enables new research and vastly enriched information dissemination.

# Year 1

## Science

### **Objectives**

Select initial science prototypes for implementation based on science interest and feasibility of implementation. Each prototype will have a design document, implementation schedule, key technical requirements, and project team. Demonstrate at the the January 2002 AAS meeting.

### **Accomplishments**

Created and demonstrated Gamma-ray burst follow-up service (a “show me the sky” service), a brown dwarf candidate search (an SDSS/ 2MASS cross-correlation service, which discovered a new brown dwarf!), and a galaxy morphology analysis (which used the Grid tools Chimera, Pegasus, and Condor to perform the analysis dynamically).

## Technology

### **Objectives**

Reach consensus on V1.0 of VOTable (XML) standard for exchange of tabular data, and develop associated I/O software libraries.

Define basic catalog and image access protocols, and utilize in science demonstrations.

Develop core metadata standards for describing astronomical datasets, and for collections of astronomical data. Plan and prototype resource directories (registries) that collect the metadata associated with data collections.

Perform initial experiments with Grid technologies to gain experience with capabilities. Evaluate and test distributed, Grid-based data storage mechanisms.

### **Accomplishments**

VOTable 1.0 released in April 2002. At least four I/O libraries developed in VO collaborations.

“Cone search” (catalog access) and Simple Image Access Protocols developed and utilized in Gamma-ray burst and galaxy morphology prototypes.

“Unified Content Descriptors” were developed for semantic description of catalog content.

Definitions were drafted for Space-Time metadata standards. Wrote initial draft of resource metadata schema for international resource registry for astronomy.

Grid software (Chimera, Pegasus, Condor) was used to support the galaxy morphology algorithm on a Grid-enabled cluster. Gained experience with the Storage Resource Broker.

## Education and Public Outreach

### **Objectives**

Gather community inputs for EPO-based requirements for the NVO. Initial focus should be on metadata needed to properly identify and classify NVO resources appropriate for EPO use cases.

### **Accomplishments**

An EPO workshop was held resulting in contacts with ~15 EPO organizations. Representative projects were identified, including the establishment of an amateur astronomy image archive with Sky & Telescope magazine. A requirements document was developed describing EPO-specific metadata.

## Year 2

### Science

#### **Objectives**

Incorporate requirements from the theoretical astrophysics community, developing a science demonstration based on theoretical simulations of globular clusters for the January 2004 AAS.

Extend Year 1 science demonstrations to usable services.

Incorporate international data access into science demonstrations.

Collect and integrate science requirements for access to spectral data. Expand data access capabilities with additional Cone Search and SIAP services.

### Technology

#### **Objectives**

Refine VOTable standard, as agreed with IVOA partners.

Re-examine the mapping of UCD structures onto data models to understand how to provide access to data collections.

Work toward international consensus on registry of astronomical resources, and on creating persistent digital identifiers.

Develop a VO Query Language standard to access the registry.

Determine strategy for integrating web services with grid technology.

Implement an NVO testbed on the NSF Teragrid, including the replication of additional image archives onto Teragrid resources (SRB). Gain experience in implementation of Web Services, including a Registry Service.

Develop initial data models for spectral and time series data, and extend SIAP interface definitions accordingly.

#### **Accomplishments**

Created development plan for science demonstration based on GRAPE globular cluster simulations.

NVO Data Inventory Service released as first NVO public service.

Continued to collaborate with IVOA to assure interoperability of US and non-US facilities. Science demonstrations for July 2003 IAU incorporated international data access.

Usage scenarios for spectral and simulation data requested and received. NED database has a NVO-compliant (SIAP) service.

#### **Accomplishments**

VOTable extensions will be considered by the IVO collaboration in Autumn 2003.

Full ontology extension of the UCD vocabulary being discussed.

Metadata attributes (VOResource) defined to support a distributed NVO registry, with two publication portals. Achieved international agreement on syntax for digital object identifiers.

VO Query Language development broken into three streams (SQL extensions, OpenSkyNode, and natural language). Role in registry queries remains under discussion.

Base current implementations on web services, and map to corresponding Grid services as they become stable.

Grid codes in development for mosaicing and multi-wavelength image federation (Atlasmaker/Montage).

Spectral data access will be through separate interface, SSAP, at least initially. Use-case scenarios being used to define protocol. Work on time series data deferred.

## Education and Public Outreach

### **Objectives**

Incorporate EPO metadata in NVO resource registry.

Update survey of important EPO resources.

Update survey of EPO-oriented access tools to NVO resources, and document for EPO users.

EPO metadata incorporated in NVO resource registry. An analysis of an interactive kiosk design for museum and planetarium partners, based on the Data Information Service. Interactions with IVOA partners will be pursued for development of common services.

Survey EPO Community projects that could most easily utilize and benefit from NVO data access.

### **Accomplishments**

EPO-related metadata integrated in resource metadata for registry.

## Year 3

By year 3, the expectation is that the Teragrid will become available for demonstrations at scale of representative NVO services. This requires both access to NVO collections and integration with grid services. The NVO service registry will be populated with services from each of the data centers.

## Science

### **Objectives**

Make direct comparisons of observed and simulated data, with focus on Globular clusters.

Execute multiple large science runs that analyze the contents of entire image archives. The Atlasmaker/ Montage package will create atlases of SDSS, 2MASS, DPOSS, FIRST, etc, and provide NVO-compliant access services.

Create wide-area atlases (digital reference sets) from multiple, large sky surveys, allowing data-mining of multi-wavelength imagery.

Port multi-parameter analysis packages to ingest NVO-compliant data services (e.g., density estimation, N-point correlation). These packages will run as services on the Grid. Plan a scientifically significant cluster/outlier search. Integrate clustering/outlier software with other packages.

Deploy core science services for use by the research community.

### **Accomplishments**

## **Technology**

### ***Objectives***

Define and develop OpenSkyNode services to provide open database access among international VO partners. Evaluate OGSA-based implementations.

Implement generalized cross-correlation services for distributed catalogs, with web and grid service support.

Complete spectral data access protocol and deploy SSAP services.

Work closely with major data providers to facilitate development of VO-compliant data access services.

Use registry services routinely for publication, discovery, and utilization of VO resources.

Prototype a knowledge engineering (ontology) approach to astronomical knowledge by extending the UCD vocabulary.

Expand NVO test-bed and test scaleability of algorithms and data access methods. Incorporate support for virtual data products.

### ***Accomplishments***

## **Education and Public Outreach**

### ***Objectives***

Promote interactions with the EPO community through the deployment of interactive kiosk, and the completion of an NVO portal for outreach.

Encourage and aid resource managers in converting data with EPO potential into EPO-ready data.

Produce a document that provides a broad overview of NVO resources for non-technical audience.

Produce a list and short description of datasets already in NVO resources that are suitable for general EPO usage without further processing.

### ***Accomplishments***

## Year 4

The NVO will plan for longer term deployment and support. Large amounts of metadata will be searchable through the NVO registry, and large amounts of data will be available through NVO-compliant standard services. Emphasis will be on supporting publication of data at all stages of creation from personal libraries to journals. Grid computing and big databases will be accessible through the Virtual Data paradigm.

We will build the 10% of services that will make 90% of customers happy!

### Science

#### **Objectives**

Search for outliers and clusters in large federated datasets via datamining algorithms.

Find faint, variable, and very extended objects in large federated datasets via datamining in the image domain.

Compare large (TB-scale) theoretical simulations with observational data.

#### **Accomplishments**

### Technology

#### **Objectives**

Increase deployment of Grid services for operating on large, federated data collections.

Provide large-scale derivation (virtual) data products, such as statistically qualified cross-matches between large surveys.

Expand registry functional to a semantic web or concept space.

Interface NVO web-based services to Grid-based counterparts.

Provide NVO development toolkit and associated training to community of researchers. VO-enabled applications developed in the community should begin to exceed those developed within the core project.

#### **Accomplishments**

### Education and Public Outreach

#### **Objectives**

Provide, with NVO partners, EPO-oriented portals to NVO data and services.

Use NVO services as a vehicle for outreach about IT itself.

Repeat and update survey of EPO community requirements and formulate plan for inclusion in NVO infrastructure.

#### **Accomplishments**

## Year 5

The NVO will transition to longer term deployment and support, including workflow systems for digital library processes for creating standard digital reference data sets and publication of scientific data. The NVO will promote full integration with Journal publishing, and wide-scale use of the NVO registry. The NVO registry will become an essential component of cyber-infrastructure for astronomy. Wide use of grid computing and big databases will be accessible through the Virtual Data paradigm.

### Science

#### **Objectives**

NVO-enabled science should be visible in the peer-reviewed literature.

NVO-based research tools will be in routine use, and will become an essential part of the environment for doing astronomical research.

#### **Accomplishments**

### Technology

#### **Objectives**

Scale registry services to larger numbers of resources and perhaps finer level of detail, with support for many simultaneous users.

Maintain core systems and improve capabilities in step with continuing evolution of underlying IT, digital library, and grid technology.

#### **Accomplishments**

### Education and Public Outreach

#### **Objectives**

Utilize NVO-compliant resources in many levels of education and outreach.

Seek interaction with NSF NSDL initiative for development of curricula based upon NVO resources.

#### **Accomplishments**